DYNAMIC RANGE AND BANDWIDTH OF ANALOG CMOS CIRCUITS

JEFFREY NORWOOD HARRISON

Bachelor of Engineering (Monash University, 1996)

A thesis submitted to Macquarie University in fulfilment of the requirements for the degree of Doctor of Philosophy

January 2002

Department of Electronics

Division of Information and Communication Sciences

Macquarie University, Sydney, Australia
CONTENTS

Abstract ................................................................. 9
Statement of originality ............................................. 11
Acknowledgement ..................................................... 12
Notation and Acronyms ............................................. 13

1 Introduction ......................................................... 17
  1.1 Dynamic range and bandwidth ................................. 17
  1.2 Aim ............................................................. 18
  1.3 Approach ....................................................... 20
    1.3.1 Analysis techniques ....................................... 20
  1.4 Scope .......................................................... 22
    1.4.1 Pitfalls of roadmapping ................................... 22
    1.4.2 Target application and frequency range ................. 23
  1.5 Synopsis ....................................................... 24

2 State-space calculation of dynamic range ....................... 29
  2.1 Introduction ................................................... 29
    2.1.1 Four questions and the organisation of this chapter .. 30
  2.2 State-space equations ......................................... 32
  2.3 Maximum signal swing and some scaling considerations .... 34
2.3.1 Maximum signal swing .............................................. 34
2.3.2 Basic points regarding power supply voltage scaling ............. 34
2.4 Noise ................................................................. 37
2.4.1 Thermal noise .................................................... 38
2.5 Examples ............................................................. 39
2.5.1 LC ladder filter .................................................. 40
2.5.2 Feedback amplifier using a Miller-compensated opamp ............ 41
2.5.3 Results ............................................................ 43
2.6 The dynamic range limitation ......................................... 46
2.6.1 Derivation ......................................................... 46
2.6.2 Interpretation ..................................................... 47
2.6.3 Application to filter dynamic range ................................ 48
2.6.4 Basic example .................................................... 48
2.7 802.11a receiver example ........................................... 49
2.7.1 Problem .......................................................... 49
2.7.2 Calculations ....................................................... 50
2.7.3 Comments ........................................................ 52
2.8 State-space realisation theory ....................................... 52
2.8.1 Controllability and observability Gramians .......................... 53
2.8.2 Similarity transformation ....................................... 54
2.8.3 Product of the Gramians ....................................... 55
2.9 An optimal realisation based on the dynamic range limitation ........ 57
2.9.1 Singular value decomposition .................................. 58
2.9.2 Numerical realisation algorithm .................................. 61
2.9.3 Properties of the optimal realisation ................................. 62
2.9.4 Example of frequency-weighted input-normal realisation ............. 65
2.10 Transfer function optimisation ..................................... 67
  2.10.1 Amplifier transfer function .................................... 68
  2.10.2 What is feedback? ............................................. 72
2.11 Links to other literature .......................................... 73
  2.11.1 The Hankel operator ......................................... 74
  2.11.2 Defining a frequency-weighted Hankel operator .................... 76
  2.11.3 Controllability and observability operators ....................... 76
  2.11.4 Optimum dynamic range realisation ............................. 77
  2.11.5 Sensitivity .................................................. 79
  2.11.6 Sensitivity theory and other related previous work .............. 80
  2.11.7 Energy storage, sensitivity and group delay .................... 81
2.12 Conclusion .................................................................. 81

3 A 350MHz opamp-RC filter ............................................. 83
  3.1 Introduction ...................................................... 83
  3.1.1 Organisation of the chapter ................................... 84
  3.2 Choice of transfer function ....................................... 84
  3.2.1 Transfer function optimisation ................................. 85
3.2.2 Graphical examination of dynamic range limitations .................. 86
3.2.3 Relationship with existing theory ......................................... 86
3.3 Realisation ................................................................. 88
3.3.1 Unsuitability of the frequency-weighted input-normal realisation .... 88
3.3.2 Merit of lossless realisations ........................................... 89
3.3.3 Lossless realisation ...................................................... 91
3.3.4 Filter taps - practical details ......................................... 92
3.3.5 Replica oscillator ....................................................... 94
3.4 Opamp design ............................................................... 96
3.4.1 Choice of opamps as the active element ................................ 96
3.4.2 Fundamental limitations on opamp bandwidth .......................... 97
3.4.3 Effect of finite opamp gain-bandwidth .................................. 98
3.4.4 Gain stage principles .................................................... 101
3.4.5 Compensation ............................................................ 104
3.4.6 Opamp circuit ............................................................. 106
3.5 Measured performance ....................................................... 106
3.5.1 Frequency response ...................................................... 106
3.5.2 Dynamic range ........................................................... 111
3.6 Additional simulations ...................................................... 112
3.7 Conclusions ................................................................. 116

4 Analytic limitations on sigma-delta modulator performance ......... 118
4.1 Introduction .................................................. 118

4.2 The sigma-delta modulator ........................................ 118
   4.2.1 Trade-off between resolution and oversampling ratio ........ 120
   4.2.2 A subjective comment regarding use of “fundamental limitations” ... 121

4.3 Noise transfer function analysis .................................. 122
   4.3.1 Quantisation noise spectrum .................................. 123
   4.3.2 Stability and maximum input signal ............................ 124
   4.3.3 Applicability of the NTF analysis to the modulator investigated .... 127

4.4 Integral limitations ............................................ 128
   4.4.1 Links with Bode’s sensitivity integral ...................... 128
   4.4.2 Interpretation ............................................. 130
   4.4.3 Derivation of integrals .................................... 136
   4.4.4 Effect of excess delay .................................. 139
   4.4.5 Excess delay with unstable loop filters ..................... 141

4.5 Connections with other theory .................................... 143
   4.5.1 Linear predictor interpretation ............................ 144
   4.5.2 Links with other system theory ............................ 146

4.6 Design implications of the integral limitations .................. 148
   4.6.1 Multiple-loop modulators .................................. 148
   4.6.2 Multiple-bit quantisers .................................. 150
   4.6.3 Compensation for excess loop delay ...................... 151

4.7 Conclusions .................................................. 154
A 520MHz sigma-delta ADC with a FIR DAC loop filter

5.1 Introduction .................................................................................................................. 155
  5.1.1 Argument for FIR DAC loop filters ................................................................. 155
  5.1.2 FIR DAC loop filter ....................................................................................... 158

5.2 Choice of noise transfer function ............................................................................. 159

5.3 Architecture details .................................................................................................... 160
  5.3.1 Error amplifier ................................................................................................. 160
  5.3.2 Latency and first-tap problems ...................................................................... 163

5.4 Circuit and layout details ............................................................................................ 164
  5.4.1 Error amplifier ................................................................................................. 164
  5.4.2 Quantiser .......................................................................................................... 167
  5.4.3 Shift register and DAC .................................................................................... 167

5.5 Test circuit board and testing .................................................................................... 168
  5.5.1 Data and clock inputs ...................................................................................... 169
  5.5.2 Digital output .................................................................................................. 170
  5.5.3 Power rails ........................................................................................................ 171
  5.5.4 Reference currents ......................................................................................... 171

5.6 Measured performance ............................................................................................... 172

5.7 Alternative architectures ............................................................................................ 174
  5.7.1 Bandpass option ............................................................................................... 176
  5.7.2 Digital filter ....................................................................................................... 176
  5.7.3 Continuous-time loop filter versus FIR DAC ................................................. 180
5.8 Sigma-delta versus Nyquist-rate converters at 20-200MHz bandwidth .................................... 181
  5.8.1 Performance comparison .......................................................... 181
  5.8.2 Limitations of sigma-delta ADC’s and Nyquist-rate ADC’s ............................. 182

5.9 Conclusion ................................................................................. 187

6 Conclusion and suggestions for future work ............................................. 189
  6.1 Conclusion ................................................................................... 189
  6.2 Suggestions for future work ............................................................. 191
    6.2.1 Circuit design ........................................................................ 191
    6.2.2 Circuit theory ........................................................................ 192

Bibliography ....................................................................................... 193
ABSTRACT

This thesis studies the dynamic range of CMOS analog circuits, with particular emphasis on the interaction between dynamic range and frequency response. Dynamic range has become a more significant issue for analog circuits recently due to the increased market for battery-powered products such as mobile telephones. The low power supply voltage of the deep-submicron CMOS processes used for system-on-a-chip integration also constrains dynamic range.

Previous authors such as Vittoz have investigated the dynamic range limitations of simple amplifier stages and the like. However, for more complicated analog circuits such as active filters and sigma-delta analog to digital converters, this is only part of the problem. The other part is how the dynamic range of a complicated circuit relates to that of the sub-blocks from which it is built.

An example familiar to many analog designers is that the dynamic range of a high-Q bandpass filter is less than that of the integrator sub-blocks from which it is built. Previous authors have been able to show a $1/Q$ dependence of the dynamic range for second-order systems. A more general relationship between the transfer function $H(j\omega)$ and dynamic range is derived here. It is based on a quantity $[H(j\omega_1) - H(j\omega_2)]/j(\omega_2 - \omega_1)$ which is a new frequency-domain representation of the so-called Hankel operator of state-space system realisation theory.

An opamp-RC filter was designed and fabricated in a deep-submicron CMOS process to assess the usefulness of this theory. The opamp-RC filter achieves state-of-the-art performance, with a remarkably high cutoff frequency, 350 MHz, and a signal swing of 0.5 $V_{p-p}$ for -40 dB distortion. Further simulations show that it should be practical to push these specifications to 800 MHz and 2 $V_{p-p}$.

Like filters, sigma-delta modulators have a non-trivial dependence between frequency response and dynamic range. An integral limitation by Bode is used to investigate the trade-off between oversampling factor and signal to noise ratio. The analysis has a common feature with the filter work, namely that the quantisation error is represented as additive noise in the same way that thermal noise in a filter is. That is, a noise transfer function
analysis is used. A new integral limitation is derived which shows the detrimental effects of excess delay in the sigma-delta modulator feedback loop. The analysis also shows that large performance gains are possible by using multiple-bit quantisers, provided that noise transfer functions with high out-of-band gain are used.

Based on the predictions of the theory, a sigma-delta modulator with a four-bit quantiser and a high out-of-band noise gain was constructed. It uses a novel topology in which the loop filtering is implemented in a finite-impulse-response digital to analog converter. This allows a sample rate of 520 Ms/s, beyond the limits of switched-capacitor technology, to be used. The fabricated sigma-delta analog to digital converter has a limited signal to noise and distortion ratio of 40dB, but achieves it over the very wide bandwidth of 60MHz. The potential of wideband sigma-delta modulators to compete with pipelined and other Nyquist-rate approaches is assessed.

The general conclusion from this work is that achievable dynamic range depends heavily on the bandwidth of the sub-blocks making up a circuit such as filter or analog to digital converter, not just their dynamic range.
STATEMENT OF ORIGINALITY

I certify that the work contained in this thesis is all my own work except where acknowledged and that the thesis has not been presented to any other University or Institution.

Candidate Signature:
ACKNOWLEDGEMENT

I express gratitude to my primary supervisor, Prof. Neil Weste. The most important of his contributions are the opportunity to work for Radiata (now part of Cisco Systems, Inc.), the provision of design tools and funding for fabrication runs which has allowed me to learn far more in four years than I would otherwise have, and non-technical skills which I have picked up by observing him. Neil’s basic message is to the effect that the shortest distance between two points is a straight line, in both design and teamwork.

I thank Cadence Design Systems, Inc. for supplying Macquarie University with the design tools used throughout this work, and Cisco Systems, Inc. for fabrication run space for the filter in chapter 3.

The ideas developed in this thesis have more distant origins. The interest in achieving high feedback over wide bandwidths essentially comes from knowing the work of Ed Cherry at Monash University, where I was an undergraduate. The appreciation of Bode’s work comes from Greg Cambrell and Ed Cherry at Monash. The work in chapter 2 has a long and tangled history, but the little-known work “Finite State Predictors for Gaussian Sequences” by Justesen (1978) provided the key observations that there is a natural state space for a given filter transfer function, and that norms of the state space might have frequency-domain expressions.

Important contributions come from friends and colleagues, notably Marie Wintrebert-Fouquet at Macquarie, Rodney Chandler and Andrew Adams at Radiata, and Katrina from Toongabbie. I thank my system administrators Rienzie Jayasekara and particularly Chris Corcoran, who have greatly increased my productivity by shielding me from irrelevant computer knowledge. Dave Barkin and the math library staff were particularly helpful during my time at Stanford.
NOTATION AND ACRONYMS

Notation

Where possible, an upper-case quantity is the Laplace transform of the lower-case
time-domain quantity eg. a system has impulse response $h(t)$ and transfer function $H(s)$.
Vector transfer functions are bold, eg. $\mathbf{F}(s)$. The list below omits symbols which are used
within a single subsection, as different symbols are often given the same name eg. $K$ or $M$.

$A^*$  conjugate transpose of $A$: $A = a_{mn} + jb_{mn} \rightarrow A^* = a_{nm} - jb_{nm}$
$A^T$  transpose of $A$
$\bar{A}$  complex conjugate of $A$
$|A|$  absolute value (magnitude) of $A$
$\|A(e^{i\theta})\|_\infty$  $\infty$-norm of the discrete-time frequency response, which is the peak gain
$\|a\|_1$  $1$-norm in the time-domain, which is $\sum_i |a_i|$
$1.44/0.24$  transistor with gate width $1.44 \mu m$ and gate length $0.24 \mu m$
$\log$  logarithm to base $e$
$\text{tr}$  trace (the sum of the diagonal entries, which equals $\sum_i \lambda_i$)
$H_{\Delta A}(s)$  transfer function $H(s)$ after $A$ is changed to $A + \Delta A$, section 2.11.5
$V_{p-p}$  volts peak to peak

Notes on matrix algebra

A useful way to think of a matrix expression such as

$$\mathbf{y} = ABCx$$

is as a block diagram in which the input signal $\mathbf{x}$ successively passes through blocks $C$, $B$
and $A$ to give the output $\mathbf{y}$. That is, in matrix notation the signal flow is from right to left.

The conjugate transpose $A^*$ is the natural generalisation of the complex conjugate of a
scalar $\bar{a}$. $AA^*$ and $A^*A$ are similar to $|a|^2$, but check whether the result is a scalar, or a
covariance matrix like those in statistical signal processing. The norm

$$\|A\|_2 = \sqrt{\text{tr}(AA^*)} = \sqrt{\text{tr}(A^*A)}$$

is analogous to $|a|$. Recall that $(AB)^* = B^*A^*$. Some
books also use $A^*$ for the adjoint of $A$, which is defined by an expression involving the inner
product. The conjugate transpose is the adjoint for a matrix, but for transfer functions it needs to be checked whether the adjoint is being taken at a spot frequency or for the operator over all frequencies.

Greek and Script Symbols

\( \mathcal{C} \) Controllability operator
\( \mathcal{O} \) Observability operator
\( \mathcal{H} \) Hankel operator
\( \lambda_i \) eigenvalues; \( \sqrt{\lambda_i(K^*W^*)} \) are the Hankel singular values
\( \Phi(s) \) input signal weight, section 2.8.1
\( \Psi(s) \) output noise weight, section 2.8.1
\( \omega \) angular frequency (rad/s)
Roman symbols

\( A, B, C, D \) matrices of state-space realisation
\( C_{gd} \) transistor gain-drain capacitance
\( C_{gs} \) transistor gate-source capacitance
\( C_{ds} \) transistor drain-source capacitance
\( e \) noise added to states
\( E \) expected value
\( E(z) \) added quantisation noise in sigma-delta modulator
\( f_{\text{max}} \) maximum frequency of oscillation
\( f_s \) sampling (clock) frequency
\( f_T \) transition frequency (the frequency where current gain \(|g_{21}|\) falls to unity)
\( F(s) \) gain to states (column vector), section 2.2
\( G(s) \) noise gain from noise added to states (row vector), section 2.2
\( g_m \) transconductance, eg. of a transistor. Also \( g_m \) in \( g_m-C \) filter
\( H(s) \) a generic transfer function
\( H(z) \) a generic discrete-time transfer function,
but also the noise transfer function in chapter 4
\( K^\Phi \) controllability Gramian, section 2.8.1
\( L(s), L(z) \) open-loop gain
\( R(z) \) sigma-delta modulator signal input
\( s \) frequency variable from Laplace transform
\( T \) a transformation of the state space
\( u \) state-space system input
\( V_{DD} \) power supply voltage
\( V_{\text{sat}} \) drain to source saturation voltage of a transistor
\( V_T \) threshold voltage of a transistor
\( W^\Phi \) observability Gramian, section 2.8.1
\( x \) state (column vector)
\( y \) state-space system output
\( z \) discrete-time frequency variable from \( Z \)-transform
Acronyms

ADC  analog to digital converter
BNC  British naval connector
CMOS complementary metal-oxide-semiconductor (the term MOS is still used despite the transistors not having metal gates)
DAC  digital to analog converter
DSP  digital signal processing
FFT  fast Fourier transform
FIR  finite impulse response
IF   intermediate frequency (in a radio)
IP3  third-order intercept point (Razavi 1998, p. 19)
I/Q  in-phase / quadrature signals
LAN  local area network
LC   inductor and capacitor
nMOS n-channel metal-oxide-semiconductor transistor
NTF  noise transfer function
pMOS p-channel metal-oxide-semiconductor transistor
OSR  oversampling ratio (Nyquist frequency / useful bandwidth)
PCB  printed circuit board
Q    quality factor (energy storage / energy loss per cycle)
QAM  quadrature amplitude modulation
ROM  read-only memory
RMS  root mean square
SAW  surface acoustic wave (passive filter)
SFDR spurii-free dynamic range (output amplitude / largest distortion spur)
SNR  signal to noise ratio
SNDR signal to noise and distortion ratio (SINAD)
SRAM static random-access memory
T/H  track and hold
1 INTRODUCTION

Dynamic range is the maximum signal a circuit can handle without overload divided by the noise or other signal processing error. Dynamic range has become a more significant analog design issue with the rise in commercial importance of battery-powered devices such as mobile telephones. This work investigates ways to optimise the dynamic range of active filters and sigma-delta analog to digital converters.

This chapter introduces the particular aspects of the dynamic range problem which were studied, and the methods used to study them. It also provides a short introduction to relevant existing literature. This is not a literature survey – existing results are introduced when required in the body of this work.

1.1 DYNAMIC RANGE AND BANDWIDTH

There seem to be two reasons why dynamic range has received attention lately. One is the desire to increase the battery life of mobile telephones, laptop computers and the like. In the past it has been possible to achieve satisfactory performance by increasing power supply current and lowering impedances to reduce the thermal noise voltage. In active filters, dynamic range could normally be made less of a problem than sensitivity. If battery life and hence power consumption is an issue this is no longer acceptable; Nokia will not buy your chip.

The other reason for a focus on dynamic range has to do with integration. The cost of a product can potentially be reduced by combining the circuitry in fewer integrated circuits. An ultimate aim is to include almost all circuitry on a single “system-on-a-chip” and to use as simple a fabrication process as possible. Now the speed of digital circuits can be increased by reducing the transistor gate length, and the power consumption can be reduced by decreasing the power supply voltage $V_{DD}$. Shorter gate length transistors require lower $V_{DD}$ to avoid gate oxide breakdown or hot electron degradation. Reducing the
power supply voltage reduces voltage swing, and hence it would be expected that analog
dynamic range would decrease. This is basically true, although the reduction in dynamic
range is less than expected as noise voltage also falls.

Fig. 1.1 shows typical transistor $V_{DD}$ over a range of process generations. We can plot
$V_{DD}$ against some other quantity; the author has chosen to plot it against a measure of
transistor speed, peak $f_T$. Given that $V_{DD}$ already decreased from $10 - 20$ V in early analog
CMOS designs to the $3.3$ V popular for analog circuits today, why are we concerned that it
will fall further? It is largely because $V_{DD}$ has fallen to the point where transistor
saturation voltage and threshold voltage are a significant fraction of power supply voltage,
requiring changes to circuit techniques. Dynamic range depends on signal swing as a
fraction of supply voltage, and this reduces if a fixed proportion of the supply voltage is lost
as a saturation or threshold voltage. The question is then whether the increased speed of
the transistors can in some way compensate. The conventional answer to this is “not
much”; the author is more optimistic.

1.2 Aim

It seems natural to try to “design-in” good dynamic range, and to do so in as systematic a
manner as possible. At the level of individual transistor stages, the limitations on dynamic
range such as maximum signal swing, distortion, thermal noise and flicker noise, are
reasonably well known. There are fundamental trade-offs between dynamic range and power
consumption; see for example (Vittoz 1994), (Castello & Gray 1985) and (Annema 1999).
However, what seems to be less clear is how the dynamic range of a system relates to that
of its sub-circuits. This is particularly true for mathematically complicated circuits such as
active filters and sigma-delta modulators. How does a sigma-delta modulator achieve higher
resolution than its quantiser? Why do narrowband bandpass filters always seem to have
poor dynamic range? This work aims to answer these questions.

A contemporary designer will usually answer the first question by saying that the
sigma-delta modulator trades off bandwidth for resolution – it uses oversampling and noise
shaping. If such a tradeoff exists between bandwidth and dynamic range (and we shall see
Figure 1.1: Power supply voltage $V_{DD}$ versus peak transition frequency $f_T$ for past, present and roadmapped processes. Labels are feature size ($\mu$m). Symbols indicate source: square – $f_T$ from long channel model in Wong & Salama (1983); circle – Kakumu et. al. (1990) with $f_T$ estimated from inverter delay; diamond – SPICE models from the foundry used by the author; triangle – predictions in Morifuji et. al. (1999).
that it does), it is reasonable to ask how much bandwidth needs to be traded off. Where is the optimum point on fig. 1.1: a deep submicron process with wide bandwidth and lower signal swings, or an older process? Obviously, the answer is likely to depend on the circuit involved and its frequency range of operation, so this work does not attempt to give a definitive answer. Rather, the aim is to develop theories of dynamic range which include bandwidth considerations.

1.3 APPROACH

This work has two basic lines of investigation:

1. Design of circuits with high dynamic range. As we want to design systems-on-a-chip in deep submicron CMOS, the circuit design concentrates on deep submicron processes.

2. Determination of fundamental limitations on the achievable dynamic range. This is in some sense the complement – designing a circuit shows that a given level of performance is achievable; finding a fundamental limitation shows that a given level of performance is not achievable.

(There is a third possibility, namely providing a non-constructive proof that a circuit with a given performance exists. The author cannot recall any such results in circuit theory.)

The author notes that the second method varies widely in appeal, and urges those readers for whom it is not a natural way of thinking to at least bear with it and see what comes from it. More will be said on this in chapter 4.

1.3.1 Analysis techniques

As dynamic range is defined by maximum signal swing and noise, we need a way to calculate the maximum internal signal swing, and the effect of internal noise on the output.
Additive noise

The theoretical work here analyses the error which limits dynamic range as an additive “noise”, whether the source of such noise is quantisation error, thermal noise or distortion. The gain from this noise to the output is a key quantity. This same approach is used for both filters (in chapter 2, using the noise gain vector $G(s)$) and sigma-delta modulators (in chapter 4, using the noise transfer function $H(z)$).

This linear analysis is widely applicable as any distortion or parameter variation can be considered as the original signal plus a “noise”. Other approaches can give more information if the “noise” has a consistent structure. Distortion cancellation and calibration become possible, and problems such as tones in sigma-delta modulators need to be considered. However, quantisation noise and particularly thermal noise are often fairly white, so the linear analysis does not miss any information. For distortion this is less true, but the author still considers that much can be learned by considering a transistor as a linear two-port with an additive drain current distortion “noise”.

Analog signal processing and causality

A distinctive feature of analog signal processing is that it is done in real time. Circuits are of course causal, so only past inputs are available. This statement seems banal, but it is important because the key derivations in chapters 2 and 4 rely on the causality property. If the output at all past and future times was available, like when running a batch digital signal processing task, the dynamic range limitations would be entirely different.

Another notable feature of analog signal processing is that storing signals is difficult. Feedback cannot be used to correct for inaccurate storage. Compare this with a digital filter where signals can be stored exactly and (rounding) error is only in multipliers and the like. This restriction, together with causality, means that a good filter or ADC works on signals in the recent past. In general, the more recent the signal the better, but at some point transistor speed becomes a limiting factor. This idea is present in the way we consider an additive noise to each integrator in chapter 2 and in the way we consider excess sigma-delta modulator loop delay in chapters 4 and 5.
1.4 SCOPE

The work tries to identify new circuit techniques. It is not a general “roadmapping” consideration of analog circuit performance versus process feature size, despite the fact that the “fundamental limitations” derived have relevance in this context. The circuit types considered are also limited.

1.4.1 Pitfalls of roadmapping

This work does not attempt to analyse in detail the effects of scaling on analog CMOS circuits. The following example helps illustrate why:

A scaling of $\lambda = 4$ would reduce the minimum channel length to $2 \mu m$ and the circuit area by approximately a factor of 16 while still keeping second-order effects within bounds. Any further scaling would lead (as will be seen in the following discussion) to unacceptable degradation in analog performance. (Wong & Salama 1983)

The “unacceptable degradation” mentioned is a combination of higher power consumption for a given dynamic range and lower opamp gain due to the increased output conductance of short-channel devices. Today, the process generation at which further scaling is considered detrimental seems to be around $0.35 \mu m$ (Annema 1999) (Buss 1999). The same reasons are given. Attempts to find a point at which scaling becomes detrimental seem to be too dependent on assumptions. To be quantitative, such analyses must use today’s circuit techniques, which are suited to today’s processes. The aim of this work is rather to identify tomorrow’s circuit techniques.

It should be noted that the general studies of analog scaling do serve the valuable function of identifying “On Growth and Form” (Thompson 1917) quality factors and figures of merit. These simplify design optimisation by making the design variables orthogonal. Some papers on analog scaling which the author has found useful are (Vittoz 1993), (Vittoz 1994), (Castello & Gray 1985), (Groenewold 1992), (Annema 1999) and (Bult 2000).
1.4.2 Target application and frequency range

This thesis concentrates on communications circuits with bandwidths of 10 MHz – 1 GHz. The Radiata (now Cisco) / Macquarie University / CSIRO 5 GHz 50 Mbit/s wireless LAN (Skellern et al. 1997) was often the target application in mind. This has several implications:

- Flicker noise is not discussed as thermal noise dominates at these bandwidths.
- The emphasis is on continuous-time circuits. Partly this is because of the clock-rate limitations of switched-capacitor circuits. Partly it is because continuous-time circuits can allow larger dynamic range because they do not “fold down” thermal noise into the passband. Partly it is because switched-capacitor circuits have already been researched by many competent researchers.
- The bandwidth is a little high to be investigating class-B and rail-to-rail opamps. “High signal swing” at these bandwidths generally means a significant fraction of rail-to-rail – the aim is to avoid circuits which are limited to swings of the order of $I_D/g_m$.

The consideration of RF CMOS was beyond the scope of this thesis. Active filters and ADC’s do actually present more of a dynamic range bottleneck than the RF signal path. VCO phase noise and digital noise coupling are big RF dynamic range issues, but they are large problems in themselves.

“Dynamic range” will mean the ratio of maximum signal to the noise at the same time. The other popular use of “dynamic range” which is connected with automatic gain control (AGC), companding and log-domain filters (that is, circuits which can have small signal and small noise at one time, and large signal and large noise at another) will not be discussed. The statement “dynamic range will degrade with CMOS scaling” is obviously not about limitations of AGC circuits.

Current-mode and switched-current circuits are not considered. Only a few state-of-the-art circuits could be designed in the time available, and the author’s guess was that feedback, voltage-mode, polysilicon resistors and routing metal capacitors were the way to get optimum dynamic range. This work tries to demonstrate that such circuits allow
performance equal to that achieved in older CMOS processes; the question of whether current-mode circuits can also achieve such performance is left open.

1.5 SYNOPSIS

The body of this work consists of a chapter of active filter dynamic range theory, a chapter describing an active filter, a chapter of sigma-delta modulator dynamic range theory, and a chapter describing a practical sigma-delta modulator, in that order.

State-space calculation of dynamic range

Chapter 2 describes how to calculate the dynamic range of a system such as an active filter, using its state-space representation. If the reader has seen the work of Groenewold (1991), they should think of the chapter as developing that paradigm. This work is to be presented as


An earlier version of this work has been presented as


However, the results in that paper are superseded by the dynamic range limitation described in chapter 2 and in (Harrison & Weste 2002).

The state-space treatment of dynamic range originates in work done on finite-wordlength effects in digital filters in the 1970’s. Storing a digital signal in a finite number of bits is clearly a limitation on the dynamic range. Both digital filters and analog active filters based on integrators have very convenient state-space descriptions. There are an infinite number of different states-space realisations for a given transfer function. The question is then which one is the best in terms of dynamic range. A reasonable answer to this was found by
Mullis & Roberts (1976) and Hwang (1977). This theory involves certain invariants known as Hankel singular values or second-order modes.

The Hankel singular values also appeared in the related work of the Russian operator theorists (Adamyan, Arov & Krein 1971). This work was taken up by the control systems community to use for model reduction – the reduction of the dimension of the state-space model of a system. Model reduction is an extreme case of finite wordlength realisation – some states are given no bits at all! Chapter 2 uses a frequency-weighted version of the dynamic range optimisation /model reduction theory, which is due to Enns (1984) and Thiele (1986). The developments in control systems and digital filtering largely proceeded independently.

The state-space theory was applied to analog active filters as it developed. The state-space description itself has a long history; see for example Anderson & Vongpanitlerd (1973). The dynamic range theory (with its Hankel singular values, Gramians, gains to states and noise gains) was only used more recently, by Snelgrove & Sedra (1986), Johns & Sedra (1987) and Groenewold (1991). The main reason for this is probably that the ‘optimal’ realisations turn out to be impractical. Because of this, the main use of the theory is for analysis and conceptual understanding. Chapter 2 continues the development in this direction. With Groenewold’s work and this work, the theory is reaching the point where it will become the 21st century textbook introduction to filter design. The first half of chapter 2 is similar to Groenewold’s and Snelgrove’s development; the second half continues it with a new limitation on dynamic range, a new frequency-domain presentation of the optimal realisation theory, and an application to feedback amplifiers.

A 350MHz opamp-RC filter

Chapter 3 presents an opamp-RC filter built in a 0.18µm standard CMOS process. The filter is notable in that it has a very high cutoff frequency (up to 350MHz) for an opamp-RC architecture, and large signal swing (0.5Vp-p differential for ~40 dB THD) considering the low $V_{DD}$ (1.8 V) and frequency of operation.

Most of the prior art – opamp-RC filters, state-space simulation of LC ladders,
feedforward and two-pole compensation of opamps – was in place by the early 1970’s. For example, see (Anderson & Vongpanitlerd 1973) and (Roberge 1975). Why research opamp-RC filters in 2001 then?

Much of the research in the intervening period involved developing active filter architectures compatible with CMOS processes. MOSFET’s were used in place of resistors to allow continuous tuning (MOSFET-C filters), and transconductor-C (gm-C) filters were developed to allow grounded MOS capacitors and supposedly to allow operation at higher frequencies. The number of papers published on transconductors is extraordinary. Switched-capacitor filters were of course also developed. However, circumstances have gradually changed to make an on-chip opamp-RC filter practical:

- Even digital CMOS processes have reasonable polysilicon resistors (non-silicided polysilicon is used in ESD protection).

- Tuning (to overcome process variations) can use resistors and CMOS switches rather than MOSFET’s. In the past, such a switched tuning made tuning loops a little more awkward. However, tuning loops are often digital today and no such problem exists.

- With the increasing number of routing metal layers, sandwich capacitors with little bottom-plate capacitance can be built in standard digital processes. Grounded capacitors are thus no longer necessary.

The final consideration is to make an opamp with sufficient bandwidth, which is the subject of much of chapter 3. Note that opamp-RC filters potentially have an advantage over gm-C filters with respect to dynamic range. This is both because feedback circuits tend to allow larger signal swing with acceptable distortion, and because the fundamental limitations on integrator dynamic range are about 4 dB better (Moreira & Silva 2001) (Groenewold 1992).

**Analytic limitations on sigma-delta modulator performance**

Chapter 4 derives fundamental limitations involving the tradeoff between sigma-delta modulator oversampling ratio and resolution. It has been presented as:

The results draw on two theories. One is the noise transfer function (NTF) linearised analysis of sigma-delta modulators. The basic approach of the NTF analysis is to regard quantisation error as an additive noise, and to calculate the noise transfer function $H(z)$ from this noise source to the output. Analysing quantisation error in this way is classical, and the NTF analysis seems to have developed over time. An important more recent aspect of this theory is that the stability of sigma-delta modulators can be roughly assessed by the maximum of the NTF $||H(e^{j\theta})||_\infty$. (This maximum of course occurs out-of-band; the aim is to make the in-band NTF small.) An example of such work is Schreier (1993). Excessive NTF’s make the sigma-delta modulator overload from its own quantisation noise.

The other theory is that of integral sensitivity limitations, which basically starts with Bode (1945). These state that while the sensitivity of a control system can be made very small at a spot frequency, it cannot be made very small over a wide bandwidth. The definition of sensitivity in control theory is the same as that of the NTF, $1/(1 + L(z))$ where $L(z)$ is the loop gain. The discrete-time version of Bode’s result and another less well known result from linear prediction theory are used here. The application of the Bode integral to sigma-delta modulators has occurred to previous authors such as Gerzon & Craven (1989); the best exposition elsewhere in the literature seems to be Nawrocki, Goldberg & Sandler (1999). (This was published after the author had re-derived and submitted similar results, although reviewers suggest that the ideas have been around for a while and were known to researchers such as Magrath and Sandler.)

The Bode integral seems to be regarded as somewhat obscure and academic by circuit designers, but it is considered one of the key results of robust control theory:

It is natural that Bode’s integral formula should have a central place in any theory of complex systems, as it was the first result to focus completely on robustness trade-offs, in this case imposed by causality. [J. C. Doyle, in (Antsaklis 2000)]
A 520MHz sigma-delta ADC with a FIR DAC loop filter

Chapter 5 presents a sigma-delta ADC using a novel architecture in which the loop filter is implemented in the DAC. The ADC clocks at 520MHz, faster than that possible with switched-capacitor techniques. This work has been presented as


The history of sigma-delta ADC’s has two eras. In early work, a wide variety of architectures were tried. Eventually the successful combination of loop filter in the forward path and switched-capacitor implementation became established. Later work concentrated on improving the switched-capacitor sigma-delta modulator, although some diversity has returned with investigation of continuous-time loop filters and mismatch shaping. The dividing line between the two eras can be taken as (Candy 1985). The work presented in chapter 5 is peculiar in that the loop filter is in the feedback path, and the implementation is not switched-capacitor (although it is discrete-time). It in fact owes a lot to earlier work such as (Tewksbury & Hallock 1978) and (Spang & Schultheiss 1962).
2 STATE-SPACE CALCULATION OF DYNAMIC RANGE

2.1 INTRODUCTION

The dynamic range of a signal-processing building block is the maximum signal which can be processed without overload, divided by the processing error. This definition can be applied to a wide variety of circuit building blocks. Indeed, it is useful to do so because similar state-space theory can be used. In this work, the idea will be used in three cases:

1. An active filter or integrator in an active filter has a dynamic range limited by the maximum signal swing, distortion and thermal noise.

2. A quantiser, be it a stand-alone ADC or the quantiser in a sigma-delta modulator, has a dynamic range limited by the number of bits. Basically we are talking about a dynamic range of $(6n + 2)$ decibels here, where $n$ is the number of bits. (Of course, comparator offsets and the like can degrade the dynamic range.)

3. A transistor stage can be considered to have a dynamic range which is the ratio of maximum signal to distortion plus noise. An output stage with a drain current which swings from $0.1\,\text{mA}$ to $0.9\,\text{mA}$ might have a total harmonic distortion around $5\%$, giving a dynamic range of $26\,\text{dB}$.

The reader may consider any one of these examples, particularly the last, an abuse of the term “dynamic range”. However, the point of this chapter (and indeed much of this thesis) is that the three can be analysed in the same way, and that it is conceptually useful to do so. The way they are analysed is to consider all noise and distortion as an additive noise on top of the correct signal. This work is limited to studying those systems which are linear apart from the additive noise. Some name for the signal processing accuracy has to be used; perhaps an alternative to dynamic range might be peak signal to noise and distortion ratio,
but this is cumbersome. The processing error will be generically termed “noise”; remember that it may be quantisation error or distortion as well as thermal noise.

It is fairly straightforward to calculate the dynamic range of sub-blocks such as opamp-RC or gm-C integrators (Groenewold 1992) (Moreira & Verhoeven 1998) (Groenewold 1991), quantisers, switched-capacitor stages (Dias, Palmisano, O’Leary & Maloberti 1992) (Castello & Gray 1985), finite-wordlength digital multipliers (Rabiner & Gold 1975) (Roberts & Mullis 1987), amplifier stages (Annema 1999) (Bult 2000) and the like. Some of these results are in this chapter. (The term “sub-blocks” is used here as a circuit such as an active filter or sigma-delta ADC is still only a building block in a complete system such as a radio receiver.) The considerations for improving the dynamic range of such sub-blocks are also well understood – for an amplifier stage, increase the transistor widths and supply current to lower the noise; for a quantiser or digital circuit, add additional bits. There are always trade-offs with power consumption.

However, this is only part of the problem. The dynamic range of an active filter is not simply the dynamic range of its integrators. For instance, a high-Q bandpass filter has less dynamic range than the integrators it is built from (Groenewold 1991) (Abidi 1992). Similarly, a negative-feedback amplifier can have more dynamic range than the stages it is built from. Consider an audio power amplifier – the class-B output stage might have 1% distortion, or a dynamic range of 40dB if distortion is counted as “noise”. However, feedback might reduce the distortion to 0.01%, or 80dB dynamic range. (The input stage of the audio power amplifier need not have 80dB dynamic range either – it does not have large output.)

2.1.1 Four questions and the organisation of this chapter

Dynamic range can be analysed by considering noise and distortion to be an additive signal. We want to analyse circuits composed of sub-blocks. An archetypal example of such a problem is an active filter where the sub-blocks are integrators. To find the dynamic range, there seem to be four natural questions:

1. What is the maximum signal swing of the sub-blocks? For the filter archetype this is
the maximum integrator output swing for acceptable distortion.

2. How large is the additive noise at the point where it is generated? We are talking here about calculating the thermal noise of an integrator, the quantisation error of a quantiser, or the noise and distortion of a transistor stage.

3. How does the signal swing in the sub-block relate to the input signal amplitude? We have restricted ourselves to systems which are linear (apart from the additive noise), so we are interested in the gain from the input to internal nodes. For active filters we will define a vector gain to states $\mathbf{F}(s)$.

4. What is the gain from this noise source to the output? For a filter, how does filter noise relate to integrator noise? We will define a vector noise gain $\mathbf{G}(s)$ from noise added to the filter states, to the output.

Section 2.3 briefly considers question 1, the maximum signal swing in analog circuits. Section 2.4 briefly considers question 2 for one noise source which ultimately limits performance, thermal noise. These problems seem relatively well understood, and the author does not have a large contribution to make. Results are mostly reviewed only to the extent that they will be used later.

The bulk of this chapter is about questions 3 and 4. To analyse a system some structure needs to be assumed. In this chapter the well-known state-space equations are used as the structure. (In chapter 4, we use another well-known structure, that of a feedback loop.) The state-space equations are introduced in section 2.2. (This section is put before the discussions of maximum signal swing and thermal noise, so that noise can be given a state-space description immediately.) Section 2.5 contains two examples of giving systems a state-space description – an LC filter and an opamp.

Section 2.6 derives a key limitation on the achievable dynamic range of a state-space system as a function of its frequency response. Section 2.7 gives an example of combining this dynamic range limitation with practical maximum signal and thermal noise results. The example is a filter needed for a particular wireless LAN architecture. Sections 2.8 and 2.9 use this result to derive a state-space filter realisation with certain optimal properties.
with respect to dynamic range. The work in these sections appears to be new. For comparison, existing related work is given at the end of the chapter in section 2.11.

The archetypal example considered in this chapter is the active filter composed of integrators. To show that the state-space theory is conceptually useful in a wider sphere, section 2.10 continues the application of it to opamps started in section 2.5.2 by discussing how feedback amplifiers are essentially “predicted” by the state-space theory.

### 2.2 STATE-SPACE EQUATIONS

A versatile formalism for studying how the dynamic range of a filter relates to the dynamic range of its integrators uses the state-space equations. This approach was first applied to digital filters (where finite-wordlength multipliers constrain dynamic range) in the 1970’s (Mullis & Roberts 1976) (Hwang 1977) (Roberts & Mullis 1987). Later it was also applied to analog active filters (Groenewold 1991) (Johns, Snelgrove & Sedra 1989). Other useful references are Thiele (1986) and Lutz & Hakimi (1988).

The state-space theory can also be usefully be applied to transistor stages. The “integrators” are the drain current of one stage flowing into the gate capacitance of the next. The fit is not quite as good as with opamp-RC or gm-C integrators or digital delays ($z^{-1}$). However, the idea allows a deeper understanding of feedback, as will be discussed later in this chapter.

The internal structure of a system (such as a filter made up of integrators) can be specified by the state-space equations

\[
\frac{dx}{dt} = Ax + Bu + e \quad (2.1)
\]
\[
y = Cx + Du \quad (2.2)
\]

where $u$ is the input, $y$ is the output, and $x$ is a vector of internal states. One element which may be less familiar is $e$, which represents noise input to the states. In an opamp-RC filter like that in the next chapter, the state $x$ is the set of integrator output voltages, and $e$ is a combination of noise current from the tap resistors and opamp noise. Anderson & Vongpanitlerd (1973) and recent undergraduate control theory textbooks such as
(Ogata 1997) or (Kuo 1995) contain further introductory material on the state-space equations. Filters not made up of integrators can also be given state-space representations using capacitor voltages and inductor currents as states.

In the Laplace-transform domain (with zero initial conditions),

\[ s\mathbf{X}(s) = A\mathbf{X}(s) + BU(s) + E(s) \]
\[ Y(s) = C\mathbf{X}(s) + DU(s), \]

which may be solved to give

\[ \mathbf{X}(s) = (sI - A)^{-1}BU(s) + (sI - A)^{-1}E(s) \]
\[ Y(s) = C\mathbf{X}(s) + DU(s) \]
\[ = (C(sI - A)^{-1}B + D)U(s) + C(sI - A)^{-1}E(s). \]

These expressions will be frequently used, so we define the quantities

\[ F(s) = (sI - A)^{-1}B \] (2.3)
\[ G(s) = C(sI - A)^{-1} \] (2.4)
\[ H(s) = C(sI - A)^{-1}B + D \] (2.5)

so that

\[ \mathbf{X}(s) = F(s)U(s) + (sI - A)^{-1}E(s) \]
\[ Y(s) = H(s)U(s) + G(s)E(s). \]

\( H(s) \) is the familiar input-output transfer function. \( F(s) \) is the gain vector from input to states. If \( F(s) \) is large, then the integrators in a filter will overload with small inputs. \( G(s) \) is the gain from the noise added to the states \( E(s) \) to the output. If \( G(s) \) is large, then the filter will have much output noise. The aim of much of this chapter is to see how small \( F(s) \) and \( G(s) \) can be made, while still realising the desired filter transfer function \( H(s) \).

A comment about the dimensions and units of some of these quantities should avoid possible confusion. \( F(s) \), \( \mathbf{x} \) and \( B \) are column vectors, assuming there is one input. \( G(s) \) and \( C \) are row vectors, assuming there is one output. \( A \) is a matrix and \( D \) is a scalar. The units are particularly unexpected. Given the symmetry between \( F(s) \) and \( G(s) \), it might be
expected that they have the same units. They do not. Suppose that the input, state and output are all voltages. \( F(s) \) maps the input voltage to the state voltages, so it is dimensionless. The transfer function \( H(s) = CF(s) + D \) is dimensionless as it is a ratio of voltages, so \( C \) and \( D \) are dimensionless. For \( (sI - A)^{-1} \) to be dimensionally consistent, \( A \) must have units of frequency like \( s \). Thus \( (sI - A)^{-1} \) has units of time. Now \( G(s) = C(sI - A)^{-1} \), so it has units of time. (\( G(s) \) is the confusing one.) The noise output \( G(s)E(s) \) is in \( V/\sqrt{\text{Hz}} \), so \( E(s) \) is in \( V\sqrt{\text{Hz}} \). (The noise \( E(s) \) is an integrator input rather than output, so it is not measured in \( V/\sqrt{\text{Hz}} \).) If some of the signals are current-mode, the units will change although the peculiar time dependences remain.

2.3 MAXIMUM SIGNAL SWING AND SOME SCALING CONSIDERATIONS

The maximum signal swing of the integrators themselves will now be discussed. The material has been covered previously by many authors (Moreira & Verhoeven 1998) (Annema 1999) (Bult 2000) (Buss 1999), so only a brief introduction is given here.

2.3.1 Maximum signal swing

Maximum signal swing is basically limited by the power supply voltage. Opamp-based and transconductor-based filters use similar amplifiers. The output stage is usually either a cascaded stage or a simple (uncascoded) common-source stage. For a cascaded output stage, the output can swing within two saturation voltages \( V_{sat} \) of each supply rail, so the swing (per side) is \( V_{DD} - 4V_{sat} \). Without cascoding, the swing is \( V_{DD} - 2V_{sat} \). It is difficult to find data on how \( V_{sat} \) has varied with scaling, but from experience it seems to have remained constant at around 0.2 V.

2.3.2 Basic points regarding power supply voltage scaling

If dynamic range was optimised by maximising signal swing, the low \( V_{DD} \) of the deep-submicron CMOS processes favoured for digital circuits (and hence systems-on-a-chip)
would be a big problem. However, the situation is not that bad.

For a given power consumption, if $V_{DD}$ halves then the power supply current can double. All impedances in the circuit can be divided by four. This halves the thermal noise voltage, and the dynamic range is unchanged. When authors (Annema 1999) calculate that dynamic range will fall with falling $V_{DD}$, they are not just claiming that signal swing will fall; they are claiming that signal swing as a percentage of the supply voltage will fall. This point is of course well known to the experts, but in casual discussion (particularly with digital designers, whose noise problems are rarely thermal) it is often missed.

Using the results in section 2.3.1, the signal swing as a fraction of $V_{DD}$ is thus approximately $1 - 0.8 \, \text{V} / V_{DD}$ for cascoded output stages and $1 - 0.4 \, \text{V} / V_{DD}$ for uncascoded output stages. Thus, without circuit changes dynamic range does degrade with scaling for common opamp and transconductor circuits. Clearly, cascodes face severe problems by the $0.13 \, \mu\text{m}$ generation ($V_{DD} = 1.2-1.5 \, \text{V}$). This trend seems to be visible today with the $0.25-0.18 \, \mu\text{m}$ generations – Miller compensated opamps with simple output stages are seen more often than they were five years ago.

For more formal derivations of these results, see (Castello & Gray 1985) for switched-capacitor filters and (Groenewold 1991), (Groenewold 1992), (Vittoz 1993) and (Vittoz 1994) for continuous-time filters. A representative simple result is that for a continuous-time class-A stage which has a maximum signal swing of $V_{DD} - \Delta V$, the minimum possible power consumption $P$ is (Annema 1999)

$$P = \frac{8\pi k T f_{\text{sig}} \text{SNR}}{1 - \Delta V / V_{DD}} \quad (2.6)$$

when outputting a single sinusoid of frequency $f_{\text{sig}}$. As discussed above, a typical $\Delta V$ is 0.8 V for cascoded stages, and 0.4 V for uncascoded stages. Note that the result is linear not quadratic with signal swing – if the signal swing is reduced, the impedance levels can also be reduced.

Using (2.6), the dynamic range loss for a $V_{DD}$ as low as 0.8 V is only 3 dB for simple common-source stages. The popular opinion seems to be that higher $V_{DD}$, say 3.3 V, is preferable to avoid this degradation. The author’s opinion is that the increased $f_T$ of deep-submicron devices will allow feedback circuits to be used to achieve signal swings
closer to rail-to-rail with acceptable distortion. The filter in chapter 3 and the sigma-delta modulator in chapter 5 are investigations of this effect. A 3 dB loss is not a huge amount to regain. It should be stressed that the aim of this work is not to argue this point, but to try to improve the dynamic range whether it gets better or worse.

There are some other scaling considerations apart from output swing, which are briefly mentioned below.

**Capacitor area**

Dividing all impedances by four multiplies all capacitances by four, so chip area may be prohibitive. This is offset by being able to run switched-capacitor circuits with higher clock rates and hence smaller capacitors for the same dynamic range (Dias, Palmisano, O’Leary & Maloberti 1992, eq. (12)). The area capacitance of metal-insulator-metal and routing metal capacitors is also greater in more modern processes (Semiconductor Industry Association 1999).

**Mismatch**

Another argument made is that the limiting “noise” is mismatch rather than thermal noise. However this has stayed roughly proportional to thermal noise (Pelgrom, Tuinhout & Vertregt 1998, fig. 4) at around $100kT$.

**Power supply noise**

If SNR is limited by power supply noise, if voltages are halved and currents are doubled, then power supply impedance has to fall by a factor of four for the same SNR. This and other signal integrity issues are beyond the scope of this thesis. Veendrick (2001) makes the interesting point that at the 0.1 μm generation, $V_{DD}$ falls below two diode drops so CMOS latch-up should not be possible as the parasitic SCR cannot latch. This allows the use of non-epi substrates without conservative design rules to prevent latchup. Non-epi substrates allow better signal isolation.
Biasing difficulties

There are also other difficulties at low $V_{DD}$. Using a common-mode voltage of $V_{DD}/2$, the input differential pair requires $V_{gs} + V_{sat} \leq V_{DD}/2$. This becomes difficult below $V_{DD} = 1.8$ V once process, temperature and $V_{DD}$ variations are taken into account. The input common-mode voltage can be moved toward $V_{DD}$ or ground in some switched-capacitor circuits however. There are also problems with CMOS switches; the switch at the output of an opamp requires $V_{DD} > 2V_T$ plus some overdrive. In a 0.18 $\mu$m process, $V_T$ is 0.5 V typical, 0.6 V at the slow corner, and close to 0.7 V at the slow corner at low temperature. $V_{DD}$ is 1.8 V nominal and perhaps $1.8$ V -10% = 1.62 V worst-case. This leaves too little switch overdrive for decent speed.

There are two process solutions for these problems (Semiconductor Industry Association 1999). One is low-threshold or zero-threshold transistors. The large foundries offer these, but do not seem to comprehend the importance of having minimum-length zero-threshold devices. The other is to use the higher voltage (often 3.3 V) digital I/O transistors for analog circuits (Buss 1999). This has merit. The possibility of using a combination of fast low-voltage transistors and slower high-voltage transistors also deserves attention. If, say, 1.8 V and 3.3 V rails are available, one option is to use 1.8 V opamps with 3.3 V nMOS-only switches.

There are also circuit design solutions. For switched-capacitor circuits, switched opamps, clock boosting and alternative switch configurations which remove the switch at the output can be used (Bult 2000). If switched tuning is used in continuous-time circuits, similar techniques are possible.

2.4 NOISE

The state-space analysis is applicable to a variety of noise sources. It was first used for digital filters, where the noise source is rounding errors due to finite wordlength, often in multipliers. For analog circuits, distortion can be included in $e$ as well as noise.

Discussion in this thesis centres on frequencies of 10 MHz to 1 GHz. As flicker noise
corners in CMOS transistors are rarely above 1 MHz (Martin, Archer, Boulin, Frei, Ng & Yan 1997), flicker noise was not a significant issue.

2.4.1 Thermal noise

For opamp-RC and switched-capacitor circuits, thermal noise is easy to calculate as opamp contributions are usually not dominant. There are two useful ways to consider thermal noise. One is to use the resistor noise voltage $\sqrt{4kT R \Delta f}$. The other is the thermodynamic consideration that for linear passive circuits there is $\frac{1}{2}kT$ of average noise energy per degree of freedom (Pécseli 2000, p. 42). This leads rapidly to the switched-capacitor noise voltage of $\sqrt{kT/C}$ using $\frac{1}{2}kT = \frac{1}{2}CV^2$. As switched-capacitor circuits are not used in this thesis (apart from one sampling switch in a less critical position in the sigma-delta modulator in chapter 5), switched-capacitor circuits will not be considered further. It would be interesting to combine the results on switched-capacitor dynamic range in (Castello & Gray 1985) with the original discrete-time version (Mullis & Roberts 1976) (Hwang 1977) of the state-space theory.

For an opamp-RC filter (fig. 2.1) with ideal (noiseless, infinite gain) opamps, the first state equation (2.1) is

$$C_f \frac{dx}{dt} + G_f x + G_i u + i_N = 0$$

$$\frac{dx}{dt} = -C_f^{-1}G_f x - C_i^{-1}G_i u - C_f^{-1}i_N$$

$$= Ax + Bu + e$$
The noise current $i_N$ over a bandwidth $\Delta f$ depends on the sum of the conductances at the summing node (Groenewold 1991) (Péceli 2000):

$$E(i_N i_N^*) = 4kT G_t \Delta f$$

The symbol * means conjugate transpose. $E()$ is expectation. The conductance $G_t$ “seen” by the summing node is a diagonal matrix with elements (Groenewold 1991)

$$g_{t(ii)} = |g_{ii}| + \sum_j |g_{f(ij)}| .$$

The absolute values are used because in differential circuits negative taps are possible by swapping negative and positive inputs. The notation here is somewhat awkward; $g_{f(ij)}$ means the $ij$th element of the $G_f$ in fig. 2.1 etc.

The state-space noise spectral density is therefore

$$\mathbf{E}(j\omega)\mathbf{E}^*(j\omega) = 4kT C_f^{-1} G_t C_f^{-*} .$$

(Of course, $C_f$ is real so $C_f^{-*} = C_f^{-T}$. Inverse and transpose commute:

$$C_f^{-T} = (C_f^T)^{-1} = (C_f^{-1})^T .$$

For approximate calculations, it is usually simpler to consider an average integrator capacitance $C_{av}$ and average total conductance at the opamp input $G_{av}$, and take the noise input at each state as

$$e_i(j\omega)_{RMS} \approx \frac{\sqrt{4kT G_{av}}}{C_{av}}$$

In addition to this, there is opamp thermal noise. However, with sensible design this is usually smaller (Groenewold 1992). For example, in the filter in the next chapter, the opamp thermal noise adds only 0.9 dB to the noise of an integrator. (For the opamp noise contribution to be small, the opamp first stage transconductance $g_{m1}$ must be large compared to the tap conductances $G_f$ and $G_i$.)

2.5 EXAMPLES

To illustrate the state-space theory, we consider two examples. One is an LC ladder filter, or an active simulation of it. Active simulations such as gyrator or leapfrog structures
inherit the state-space equations of the LC prototype they were based on (Johns et al. 1989). Such “lossless” structures are known to have low sensitivity and high dynamic range (Orchard, Temes & Cataltepe 1985). The other example is an amplifier stage using a Miller-compensated opamp. This is a less traditional use of the state-space theory. It is important because much of this thesis is about how feedback affects dynamic range. It is also a step towards the author’s (currently unrealised) goal of synthesising opamps in the same way that filters are synthesised.

2.5.1 LC ladder filter

The state-space equations for an LC ladder filter are readily derived using Kirchoff’s laws and the constitutive relations (Anderson & Vongpanitlerd 1973, p. 147). The states are taken as the capacitor voltages and inductor currents:

\[
\mathbf{x} = \begin{bmatrix} v_{C_1} & i_{L_2} & v_{C_3} & i_{L_4} & v_{C_5} \end{bmatrix}^T
\]
The coefficient matrices are:

\[
A = \begin{bmatrix}
-\frac{1}{C_1} & -\frac{1}{C_5} & 0 & 0 & 0 \\
\frac{1}{L_2} & 0 & -\frac{1}{L_3} & 0 & 0 \\
0 & \frac{1}{C_5} & 0 & -\frac{1}{C_3} & 0 \\
0 & 0 & \frac{1}{L_4} & 0 & -\frac{1}{L_4} \\
0 & 0 & 0 & \frac{1}{C_5} & -\frac{1}{C_5}
\end{bmatrix}
\]

\[
B = \begin{bmatrix}
\frac{2}{C_1} \\
0 \\
0 \\
0 \\
0
\end{bmatrix}
\]

\[
C = \begin{bmatrix}
1 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

\[
D = \begin{bmatrix}
0
\end{bmatrix}
\]

For the Butterworth response with 1 rad/s -3dB frequency used, \(C_1 = C_5 = 0.618\), \(L_2 = L_4 = 1.618\) and \(C_3 = 2\).

### 2.5.2 Feedback amplifier using a Miller-compensated opamp

The circuit investigated is shown in fig. 2.3. This might typically be used as an intermediate frequency or baseband amplifier in a radio receiver. \(C_1\) and \(C_2\) represent transistor \(C_{gs}\) of the first and second stages (and \(C_{ds}\) and parasitics capacitances etc.). The small-signal equivalent circuit is for the differential-mode signal.

Transistor \(C_{gd}\) is omitted as it complicates the analysis. For the first stage, \(g_{m1}\) is reversed in sign to represent the left-hand-side opamp input being connected to the right-hand-side opamp output and vice versa. This is difficult to analyse if \(C_{gd}\) is present. There are also difficulties relating to writing state equations for loops of capacitors. It can be done using nodal analysis and converting \((sC + G)\) into \(C(sI + C^{-1}G)\), but the results are more difficult to interpret as a state no longer represent the voltage of a single capacitor.

The state-space equations were written using the states \(x_1\) to \(x_4\) as labelled on fig. 2.3. The component values used were:
Figure 2.3: Feedback amplifier, opamp used, and small-signal equivalent circuit for the state-space analysis.
<table>
<thead>
<tr>
<th>Component</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>5kΩ</td>
<td>input resistor</td>
</tr>
<tr>
<td>R5</td>
<td>15kΩ</td>
<td>feedback resistor</td>
</tr>
<tr>
<td>R3</td>
<td>5kΩ</td>
<td>load resistance and $g_{ds}$ of output stage</td>
</tr>
<tr>
<td>C3</td>
<td>200µF</td>
<td>load, routing, second stage $C_{ds}$, C4 bottom-plate</td>
</tr>
<tr>
<td>gm1</td>
<td>0.7mS</td>
<td>first stage</td>
</tr>
<tr>
<td>gm2</td>
<td>2mS</td>
<td>second stage</td>
</tr>
<tr>
<td>C1</td>
<td>80µF</td>
<td>$C_{gs}$ of first stage, routing parasitics</td>
</tr>
<tr>
<td>C2</td>
<td>80µF</td>
<td>$C_{gs}$ of second stage and $C_{ds}$ of first stage</td>
</tr>
<tr>
<td>C4</td>
<td>100µF</td>
<td>Miller compensation</td>
</tr>
<tr>
<td>R4</td>
<td>1kΩ</td>
<td>removes Miller comp. right-half-plane zero</td>
</tr>
</tbody>
</table>

The values used are fairly realistic for a wideband IF amplifier stage. The transistors are 18/0.36 for the input stage and 18/0.18 for the output stage. The input stage uses around 100µA per transistor and the output stage uses around 250µA per transistor. The load is a similar amplifier stage. The stage gain is 3 (9.5 dB).

Recall that the noise gain $G(j\omega)$ is not dimensionless. For an otherwise identical system, if the bandwidth is doubled, then the noise gains halve. To allow the feedback amplifier to be compared with the LC ladder filter, all of the capacitances in the feedback amplifier were multiplied by a factor of $2.5 \times 10^3$, to bring the feedback amplifier bandwidth down from 400MHz to 1 rad/s.

### 2.5.3 Results

Fig. 2.4 shows the transfer function $H(j\omega) = [h_1]$, gains to states $F(j\omega) = [f_1 f_2 \cdots]^T$, and noise gains $G(j\omega) = [g_1 g_2 \cdots]$ for both the LC ladder and feedback amplifier examples.

**LC filter**

All elements of $F(j\omega)$ are 0 dB at DC, as the current through the inductors is unity and the voltage across the capacitors is unity. The noise gains $G(j\omega)$ have different DC values
Figure 2.4: Transfer function $H(j\omega)$, gain to states $F(j\omega)$ and noise gains $G(j\omega)$ of the LC filter (left) and opamp (right) examples.
because they depend on the component value.

The signal levels here are unscaled; if this filter was used in a radio receiver it would be desirable to scale down $f_1$ to prevent the first state overloading. This would increase $g_1$ by the same factor. The scaling depends on the application; for instance the unscaled levels here would be good for a transmit filter designed to handle energy mostly concentrated in the passband.

**Feedback amplifier**

Think of the elements of $F(j\omega)$ as the AC voltages in the circuit for 1 V AC input. $f_1$ is the error voltage, which is small at low frequencies where the opamp gain is high. $f_2$ is the gate voltage of the second stage which is also small at low frequencies. $f_3$ is the output, and is therefore equal to $H(j\omega)$. $f_4$ is the voltage across the compensation capacitor, which is similar to the output.

The noise gains $G(j\omega)$ also make sense. The highest gain is from noise injected into the opamp input, $g_1$. Fortunately transistor distortion currents do not flow into this node. The noise gain from the output of the first stage is $g_2$, and is somewhat lower. Distortion currents from the first stage flow into this node, so it is fortunate that the signal swings associated with the first stage ($f_1$ and $f_2$) are small and hence should have low distortion component. The noise gain $g_3$ from distortion currents injected into the output node is very small at low frequencies. This is the feedback working – it is important that $g_3$ be small as the output stage is likely to generate significant distortion current. $g_4$ represents the gain from noise in the compensation capacitor, and is quite large. Being a passive component, hopefully the compensation capacitor should generate little noise and distortion.

In the introduction (section 2.1) it was claimed that the components in a feedback amplifier can have less dynamic range than that required of the output. It can be verified that $|f_1||g_1|$, $|f_2||g_2|$ and $|f_3||g_3|$ are all significantly less than one at low frequencies where the feedback “works”. The LC filter does not show this behaviour. Thus the dynamic range required of the components really is reduced by a feedback amplifier.

However, note that $|f_4||g_4|$ in the feedback amplifier is large. Thus the compensation
capacitor (whose voltage is state $x_4$) requires high dynamic range. This should be achievable by using a linear capacitor such as a metal-insulator-metal or routing metal structure.

Do circuits exist in which all states do not require high dynamic range? The theory which follows provides some information.

2.6 THE DYNAMIC RANGE LIMITATION

This section derives a key relationship, (2.8), which the author calls the “dynamic range limitation”. This name is used as the result limits the filter dynamic range achievable with a given integrator dynamic range and filter transfer function.

2.6.1 Derivation

The following identity will be used:

$$(s_1 I - A)^{-1} - (s_2 I - A)^{-1} = (s_2 - s_1)(s_1 I - A)^{-1}(s_2 I - A)^{-1}$$

The proof is straightforward:

$$M = (s_1 I - A)^{-1} - (s_2 I - A)^{-1}$$

$$(s_1 I - A)M(s_2 I - A) = (s_2 I - A) - (s_1 I - A)$$

$$(s_1 I - A)M(s_2 I - A) = (s_2 - s_1)I$$

$$M = (s_2 - s_1)(s_1 I - A)^{-1}(s_2 I - A)^{-1}$$

This is known as the resolvent identity (the expression $(sI - A)^{-1}$ is called the resolvent). Here $s_1$ and $s_2$ are scalars; think of them as two frequencies.

Now using $M$ in the above proof,

$$CMB = C(s_1 I - A)^{-1}B - C(s_2 I - A)^{-1}B$$

$$= C(s_1 I - A)^{-1}B + D - C(s_2 I - A)^{-1}B - D$$

$$= H(s_1) - H(s_2)$$
but also

\[
CMB = (s_2 - s_1)C(s_1 I - A)^{-1}(s_2 I - A)^{-1}B \\
= (s_2 - s_1)G(s_1)F(s_2)
\]

so

\[
\frac{H(s_1) - H(s_2)}{s_2 - s_1} = G(s_1)F(s_2) \tag{2.8}
\]

Eq. (2.8) will be called the “dynamic range limitation” in the discussion which follows. It relates the transfer function \(H\) at two frequencies \(s_1\) and \(s_2\) to the inner product of the noise gain \(G(s_1)\) at one frequency and the gain to states \(F(s_2)\) at the other.

### 2.6.2 Interpretation

The dynamic range limitation (2.8) states that if the transfer function changes rapidly with frequency, then at least one state must have a sizable product of the gain to state and noise gain. That is, at least one state must require sizable dynamic range. The intuitive explanation for why this is necessary is that if \(H(s)\) varies with frequency then past signals must be stored, and they must be stored with dynamic range similar to that required of the output.

The quantity \(G(s_1)F(s_2)\) is a map from the \(s\)-plane to the \(H(s)\)-plane. If two points \(s_1\) and \(s_2\) are close together in the \(s\)-plane, but far apart in the \(H(s)\)-plane, then \(G(s_1)F(s_2)\) has to be large. The map acts on the set of chords rather than on individual points.

If a filter was built from resonators with a transfer function \(1/K(s)\) rather than integrators with a function \(1/s\), then the quantity in the dynamic range limitation would become \([H(s_1) - H(s_2)]/[K(s_2) - K(s_1)]\). Thus the best possible dynamic range depends on the relative quality of the filter and the resonators it is built from. The Leeson (1966) equation for oscillator phase noise is also a result of this nature.
2.6.3 Application to filter dynamic range

A radio receiver is an archetypal example of a circuit requiring high dynamic range. Usually the greatest problem is to receive a small in-band signal in the face of large out-of-band interferers.

Let \( s_1 = j\omega_P \) be a frequency in the receiver passband, and \( s_2 = j\omega_S \) be the frequency of a large stopband interferer. Assume that \( |H(j\omega_P)| > k_P \), and \( |H(j\omega_S)| < k_S \). Then

\[
\left| \frac{H(j\omega_P) - H(j\omega_S)}{j(\omega_S - \omega_P)} \right| \geq \left| \frac{k_P - k_S}{\omega_S - \omega_P} \right|
\]

so

\[
|G(j\omega_P)F(j\omega_S)| \geq \left| \frac{k_P - k_S}{\omega_S - \omega_P} \right|
\]

or

\[
\sum_i |f_i(j\omega_S)||g_i(j\omega_P)| \geq \left| \frac{k_P - k_S}{\omega_S - \omega_P} \right| \tag{2.9}
\]

If the input signal is large at the blocking frequency \( \omega_S \), it is important to have small gains to states \( f_i(j\omega_S) \) to prevent the states overloading. It is also important to have low output noise in the passband, which requires small noise gains \( g_i(j\omega_P) \). Eq. (2.9) shows that this is only possible if there is sufficient filter transition bandwidth \( |\omega_S - \omega_P| \). \( |k_P - k_S| \) cannot be changed without more passband droop or much less stopband attenuation.)

2.6.4 Basic example

The following example gives an indication of whether (2.9) is tight enough to be useful.

We use the 5th order Butterworth filter analysed earlier in this chapter. It is \(-3\) dB at 1 rad/s, and \(-40\) dB at 2.5 rad/s. We will take these as \( \omega_P \) and \( \omega_S \) respectively. The stopband gains to states and passband noise gains are as follows:
\[
\begin{array}{ccc}
  i & f_i(j\omega_s) & g_i(j\omega_P) & f_i(j\omega_s)g_i(j\omega_P) \\
  1 & 1.229\angle-52^\circ & 0.219\angle135^\circ & 0.269\angle173^\circ \\
  2 & 0.321\angle-142^\circ & 0.673\angle167^\circ & 0.216\angle51^\circ \\
  3 & 0.068\angle128^\circ & 1.144\angle-135^\circ & 0.078\angle-97^\circ \\
  4 & 0.019\angle43^\circ & 1.327\angle-60^\circ & 0.025\angle103^\circ \\
  5 & 0.010\angle-14^\circ & 0.489\angle18^\circ & 0.005\angle-33^\circ \\
\end{array}
\]

This results in

\[
\sum_i f_i(j\omega_s)g_i(j\omega_P) = 0.477\angle45^\circ \\
\frac{H(j\omega_P) - H(j\omega_s)}{f(\omega_S - \omega_P)} = 0.477\angle45^\circ \\
\sum_i \left| f_i(j\omega_s) \right| \left| g_i(j\omega_P) \right| = 0.592 \\
\frac{k_P - k_S}{\omega_S - \omega_P} = 0.465
\]

As expected, the dynamic range limitation (2.8) is obeyed. What is more remarkable is how close (2.9) is to equality. This illustrates the well-known fact (Orchard et al. 1985) that LC ladders and active simulations thereof are “good” filters with close to optimum dynamic range.

2.7 802.11A RECEIVER EXAMPLE

The following example puts together the results derived so far.

2.7.1 Problem

An 802.11a 5 GHz wireless LAN must receive a −65dBm 54Mbit/s signal in the face of a −50dBm alternate adjacent blocking signal (IEEE Std. 802.11a 1999). The alternate adjacent channel starts 20 MHz from the passband edge.

If a superheterodyne 5 GHz receiver is used, the intermediate frequency (IF) must be at least, say, 300 MHz for the image to be sufficiently far away to filter out with a SAW, LC or
microstrip filter. The next chapter shows that on-chip filtering at 300 MHz is certainly possible, although the Q-tuning required for 20 MHz passband centred at 300 MHz would be delicate. Is such a filter practical from a dynamic range perspective?

2.7.2 Calculations

Filter dynamic range

The 54Mbit/s signal is multi-tone 64-QAM, which with white noise requires a signal to noise ratio (SNR) of around 20 dB for 10⁻⁶ bit error rate. As there are coding losses and are other noise sources such as the DAC, ADC and digital filters, it seems sensible to aim for an SNR of at least 30 dB for the filter.

The blocking signal is −50 dBm, but multi-tone modulation produces a high peak-to-average-power ratio. Some margin for this is needed, say that overload should not occur below −44 dBm. The filter dynamic range therefore needs to be 30 dB + (−44 dB − −65 dB) = 51 dB.

Signal swing

The best signal swing which could be hoped for is the $V_{DD} - 2V_{sat}$ per side mentioned in section 2.3.1 for an unascoded common-source output stage. In a 0.18 μm process this is typically 1.4 $V_{p-p}$ per side, or 2.8 $V_{p-p}$ differential. In practice, at high frequencies it is very difficult to achieve this with acceptable distortion. The filter in the next chapter achieves −40 dB distortion at 0.5 $V_{p-p}$ differential signal swing. The extensions in section 3.6 show that 2 $V_{p-p}$ appears achievable. We will assume a 1 $V_{p-p}$ maximum differential signal swing (0.25 V peak per side).

Filter components

Assume that the power budget is around 20 mW. Assume there are three pairs of integrators; 2–4 would be typical. This is 3.3 mW per opamp, or 1.9 mA for $V_{DD} = 1.8$ V. If the output stage current is 0.6 mA per side, a current swing of 0.4 mA peak per side seems achievable. The voltage swing is 0.25 V peak per side, so the load on an opamp should be
0.25 V / 0.4 mA = 625 Ω. This load is typically one capacitor and two or three resistors, so a unit resistor should be around 2 kΩ.

To achieve a 300 MHz pole frequency with a 2 kΩ resistor, the integrator capacitors need to be 250 fF.

**Dynamic range limitation**

Assume that the input, state and output signals are similar in magnitude. Then the gains to states $f_i(jω_S)$ should not exceed 0 dB in magnitude. Assume that $|H(jω_P)| = 0.9$ (−1 dB) and $|H(jω_S)| = 0.03$ (−30 dB). Then from (2.9),

$$\sum_i |f_i(jω_S)||g_i(jω_P)| \geq \left| \frac{k_P - k_S}{ω_S - ω_P} \right|$$
$$\sum_i |g_i(jω_P)| \geq \frac{0.9 - 0.03}{2π \times 20 \times 10^6 \text{ rad/s}} = 6.9 \times 10^{-9} \text{s}$$

**Integrator noise**

Using (2.7), the noise input is

$$E_i(jω)_{RMS} \simeq \frac{\sqrt{4kT G_{av}}}{G_{av}} = \frac{\sqrt{4 \times 1.38 \times 10^{-23} \times 300 \times 1 \times 10^{-3}}}{250 \times 10^{-15}} = 16 \text{ V}\sqrt{\text{Hz}}$$

assuming that the average conductance $G_{av}$ connected to each opamp summing node is 1 mS. This is per side; the differential noise is $\sqrt{2} \times 16 \text{ V}\sqrt{\text{Hz}} = 23 \text{ V}\sqrt{\text{Hz}}$.

**Filter noise**

As output noise is proportional to $\sum_i |g_i|^2$ but the dynamic range limitation constrains $\sum_i |g_i|$, minimum noise occurs if equal noise is produced by all six poles. As

$$\sum_i |g_i(ω_P)| = 6.9 \times 10^{-9} \text{s}, \text{ each } |g_i(ω_P)| = 1.15 \times 10^{-9} \text{s}. \text{ In this case, the output noise contribution from each integrator is}$$

$$1.15 \times 10^{-9} \text{s} \times 23 \text{ V}\sqrt{\text{Hz}} = 26 \text{ nV}/\sqrt{\text{Hz}}$$

Hence the total noise is

$$26 \text{ nV}/\sqrt{\text{Hz}} \times \sqrt{6} = 65 \text{ nV}/\sqrt{\text{Hz}}$$
Over a 20 MHz bandwidth, this integrates to 290 μV RMS.

In practice, the dynamic range is limited by the first pair of integrators and the noise is not so evenly averaged over the six poles. Based on the example in section 2.5.1 and actual calculations, active simulations of LC ladders are likely to achieve around 700 μV RMS noise. The differential signal swing is 0.35 V RMS, so the dynamic range achieved is 54 dB. This is just above the minimum required.

2.7.3 Comments

Clearly such an architecture is marginal from a dynamic range perspective. Only thermal noise from the resistors has been considered; on top of this is distortion and noise from the opamps. Against this, a signal swing of 2Vp-p differential might be possible, and the power budget could be increased a little. (There is a lot of broadband digital signal processing in a 54 Mbit/s wireless LAN, so 20 mW is unlikely to be a large component of the total power consumption.)

To achieve higher dynamic range, a Q-enhanced LC filter could be used. This cuts the power consumption for a given integrator dynamic range, as energy is stored for multiple cycles rather than being taken from the power supply each cycle. The thermal noise currents can also be lower. However, the frequency variation of inductor loss needs to be carefully considered when designing the Q tuning, and the negative resistance to cancel the losses still has to be linear for swings of a significant fraction of rail-to-rail.

2.8 STATE-SPACE REALISATION THEORY

The dynamic range limitation is useful in its own right. However, what is surprising is that it can also be used to construct filter realisations with dynamic range which is in some sense optimum. These "optimal" realisations were first found in the 1970’s and 1980’s using a time-domain and state-space approach. However, the new frequency-domain approach presented here seems simpler and more comprehensible. (It should be noted that the original algorithms are more computationally efficient though.)
In state-space terms, “realisation” means finding a set of matrices \((A, B, C, D)\) which give the correct transfer function \(H(s) = C(sI - A)^{-1}B + D\). The key problem is that there are different sets \((A, B, C, D)\) which realise the same transfer function. Moreover, these have different dynamic range properties, as the gains to states \(F(s)\) and noise gains \(G(s)\) are different. The problem we are trying to solve is to find a good set. Section 2.8.2 considers what sets exist, and section 2.8.1 considers what is a good one.

### 2.8.1 Controllability and observability Gramians

It turns out to be important to consider the gains to states \(F(s)\) and noise gains \(G(s)\) over all frequency rather than at spot frequencies. Enns (1984) and Thiele (1986) have defined the frequency-weighted controllability Gramian matrix

\[
K^\Phi = \frac{1}{2\pi} \int_{-\infty}^{\infty} [F(j\omega)\Phi(j\omega)][F(j\omega)\Phi(j\omega)]^* d\omega
\]

and the frequency-weighted observability Gramian matrix

\[
W^\Psi = \frac{1}{2\pi} \int_{-\infty}^{\infty} [\Psi(j\omega)G(j\omega)]^* [\Psi(j\omega)G(j\omega)] d\omega.
\]

(The symbol * is conjugate transpose.) The way to understand what these quantities are is to look at their definitions. \(K^\Phi\) is a weighted integral over frequency of the power gain to states. The clearest way to think of this is that the diagonal element \(k^\Phi_{ii}\) of \(K^\Phi\) is the signal power in the \(i\)th state for an input signal \(\Phi(s)\). The off-diagonal terms \(k^\Phi_{ij}\) represent correlations between the signals in the states. \((K^\Phi\) and \(W^\Psi\) are both \(n \times n\) matrices for a system with \(n\) states.) Similarly, \(W^\Psi\) is a weighted integral of the noise power gain. The noise output is weighted by an output filter \(\Psi(s)\). The diagonal elements \(w^\Psi_{ii}\) represent weighted output noise power gains, and the off-diagonal elements \(w^\Psi_{ij}\) represent frequency-weighted correlations between the noise gains from different noise inputs \(e_i\) and \(e_j\).

To optimise the dynamic range for input signal \(\Phi(s)\) and output noise weighting \(\Psi(s)\), we want to make the \(k^\Phi_{ii}\) and \(w^\Psi_{ii}\) small. A brute-force way of doing this would be to numerical minimise \(\sum_{i=1}^{n} k^\Phi_{ii}w^\Psi_{ii}\) over the space of all realisations \((A, B, C, D)\) which give a
transfer function $H(s) = C(sI - A)^{-1}B + D$ which meets the passband ripple and stopband attenuation specifications. The main problem with doing this is that the optimisation problem is too hard. For a fifth-order filter, $(A, B, C, D)$ has 36 parameters. This is quite practical for a convex optimisation, but the passband ripple and stopband attenuation specifications appear to give a non-convex problem. Nevertheless, this approach should not be ruled out, as a careful choice of a sparse $(A, B, C, D)$ could reduce the dimensionality. For example, $(A, B, C, D)$ could correspond to the elements of an LC ladder. The optimisation could also tweak an existing realisation within a convex region. The other reason why such an approach is less theoretically interesting is that it offers very little design insight.

Because of these problems, the most successful approach to filter design has been to split the problem into choosing a transfer function and choosing a realisation of that transfer function.

### 2.8.2 Similarity transformation

Different realisations $(A, B, C, D)$ can be parameterised by substituting a transformed state

$$v = Tx$$

so that

$$\frac{dx}{dt} = Ax + Bu$$

$$T^{-1}\frac{dv}{dt} = AT^{-1}v + Bu$$

$$\frac{dv}{dt} = TAT^{-1}v + TBu$$

and

$$y = Cx + Du$$

$$y = CT^{-1}v + Du.$$
The gain to states and noise gain transform as follows:

\[
F_v(j\omega) = (sI - TA^{-1})^{-1}TB = T(sI - A)^{-1}B = TF_x(j\omega) \quad (2.12)
\]

\[
G_v(j\omega) = CT^{-1}(sI - TA^{-1})^{-1} = C(sI - A)^{-1}T^{-1} = G_x(j\omega)T^{-1} \quad (2.13)
\]

Here \(F_v(j\omega)\) and \(G_v(j\omega)\) are the gain to states and noise gain for the transformed state variable \(v\), and \(F_x(j\omega)\) and \(G_x(j\omega)\) are the gain to states and noise gain for the original state variable \(x\).

Note that

\[
G_v(j\omega_1)F_v(j\omega_2) = G_x(j\omega_1)T^{-1}TF_x(j\omega_2) = G_x(j\omega_1)F_x(j\omega_2).
\]

If this did not hold, the dynamic range limitation could not be true.

The main significance of the similarity transformation is that it describes all minimal state-space realisations with a given transfer function (Anderson & Vongpanitlerd 1973, pp. 101–103). (“Minimal” basically means having the same number of integrators as the order of the transfer function. There is no evidence that non-minimal realisations allow superior dynamic range, although it has not been proved impossible.)

2.8.3 Product of the Gramians

We could go on to calculate how \(K^\phi\) and \(W^\psi\) transform under a similarity transformation. The answer is not particularly interesting; they do not stay the same. What is more interesting is how the product \(K^\phi W^\psi\) transforms. From the definitions (2.10) and (2.11) of
the frequency-weighted Gramians,

\[
K_{\Psi}^{\Phi}W_{\Psi}\Psi
= \frac{1}{2\pi} \int_{-\infty}^{\infty} [F_\Psi(j\omega)\Phi(j\omega)]F_\Psi(j\omega)\Phi(j\omega)^T \, \mathrm{d}\omega \cdot \frac{1}{2\pi} \int_{-\infty}^{\infty} [\Psi(j\omega)G_\Psi(j\omega)]^*\Psi(j\omega)G_\Psi(j\omega) \, \mathrm{d}\omega
= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F_\Psi(j\omega_1)\Phi(j\omega_2)]F_\Psi(j\omega_2)\Phi(j\omega_1)^* \Psi(j\omega_1)G_\Psi(j\omega_1)^*\Psi(j\omega_1)G_\Psi(j\omega_1) \, \mathrm{d}\omega_1 \mathrm{d}\omega_2
= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_\Psi(j\omega_1)\Phi(j\omega_2)^* \Psi(j\omega_1)F_\Psi(j\omega_2)\Phi(j\omega_2)^* \Psi(j\omega_1)G_\Psi(j\omega_1)^*\Psi(j\omega_1)G_\Psi(j\omega_1) \, \mathrm{d}\omega_1 \mathrm{d}\omega_2
= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T F_\Psi(j\omega_1)\Phi(j\omega_2)^* \Psi(j\omega_1)F_\Psi(j\omega_2)\Phi(j\omega_2)^* \Psi(j\omega_1)G_\Psi(j\omega_1)^*\Psi(j\omega_1)G_\Psi(j\omega_1)T^{-1} \, \mathrm{d}\omega_1 \mathrm{d}\omega_2
= TF_\Psi^{\Phi}W_\Psi T^{-1}.
\]

For any non-singular transformation \( M \), the eigenvalues of \( TMT^{-1} \) are the same as those of \( M \). (As an example, recall that \( A \) also becomes \( TAT^{-1} \) under a similarity transform. We know that this leaves the transfer function unchanged, so the poles are unchanged. The poles of the system are the eigenvalues of \( A \).) Thus the eigenvalues of \( K_\Psi^{\Phi}W_\Psi \) are invariant under a similarity transformation, and hence are the same for all minimal realisations of a given transfer function. These eigenvalues \( \lambda_i(K_\Psi^{\Phi}W_\Psi) \) are important, and (their square roots) are called frequency-weighted Hankel singular values.

The reason for this name will be apparent by the end of the chapter.

The Hankel singular values \( \sqrt{\lambda_i(K_\Psi^{\Phi}W_\Psi)} \) are important because they restrict how small the powers in the states \( k_i^\Phi \) and the noise power gains \( w_i^\Psi \) can be made. The matrices \( K_\Psi^{\Phi} \) and \( W_\Psi \) cannot be both small if \( K_\Psi^{\Phi}W_\Psi \) is big. The diagonal elements \( k_i^\Phi \) and \( w_i^\Psi \) cannot be small if \( K_\Psi^{\Phi} \) or \( W_\Psi \) are big, because like other covariance matrices, \( K_\Psi^{\Phi} \) and \( W_\Psi \) are positive-definite (and symmetric). There is rigorous theory concerning which norms to define “big” and “small” in (Mullis & Roberts 1976) (Hwang 1977). The results become somewhat messy because the norms are not quite the ones we would like, so we restrict ourselves to an intuitive understanding.
2.9 AN OPTIMAL REALISATION BASED ON THE DYNAMIC RANGE LIMITATION

The difference between a good realisation of a filter transfer function and a bad realisation seems somewhat mysterious at first. By a good realisation we mean one with high dynamic range and low sensitivity to component variations – fortunately these two properties are compatible. Actually, it is reasonably straight-forward. A good realisation stores signals in its states which are as orthogonal as possible, and it stores signals which are relevant to the output. The situation is similar to conditioning of matrices – a transformation which diagonalises the matrix improves conditioning as results are no longer produced by subtracting large, nearly equal quantities. A good realisation “diagonalises” the gains to states and noise gains. The Gramians $K^{\phi}$ and $W^{\psi}$ become diagonal or diagonally dominant. In terms of the discussion in the previous section, this is the way to minimise the powers in the states $k_{ii}^{\phi}$ and weighted noise output $w_{ii}^{\phi}$ for a given transfer function and hence given Hankel singular values.

The first procedures for doing this were discovered in the mid-1970’s (Mullis & Roberts 1976) (Hwang 1977). These involved finding a similarity transformation $T$ which simultaneously diagonalises $K^{\phi}$ and $W^{\psi}$. (This work was done for digital filters, which are discrete-time. However, the continuous-time theory is essentially the same.) Independently, mathematicians developed similar techniques (Adamyan et al. 1971). The original work was for weightings $\Phi(s) = \Psi(s) = I$, but it was generalised to the frequency-weighted case by Enns (1984) and Thiele (1986).

A new frequency-domain solution to the problem is presented here. It is a direct approach, in contrast to most current techniques which are based on transforming an existing realisation. Direct approaches do already exist, developed by Zeiger & McEwen (1974) and Kung (1978) from the work of Ho & Kalman (1966). However, they seem not to have been extended to the frequency-weighted case.

The frequency-domain solution shows the significance of the dynamic range limitation. However, it should be noted that the transformation-based approaches are often more computationally efficient, so the best approach may be to understand the frequency-domain
solution but to use the traditional algorithms (Enns 1984) (Thiele 1986) (Anderson & Moore 1989, problem 10.2-1). Previous approaches are discussed in detail in section 2.11.4 at the end of the chapter.

2.9.1 Singular value decomposition

The singular value decomposition of a matrix $K$ is (Golub & Van Loan 1996, p. 69)

$$K = USV^*,$$

where $U$ and $V$ are unitary matrices and $S$ is a diagonal matrix with real, non-negative entries $\sigma_i$. The $\sigma_i$ are called singular values. They are almost always ordered from largest to smallest, $\sigma_1 > \sigma_2 > \cdots$, and the matrix $S$ is arranged as

$$S = \begin{bmatrix} \sigma_1 & 0 & 0 & \cdots \\ 0 & \sigma_2 & 0 & \cdots \\ 0 & 0 & \sigma_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

A unitary matrix is like a rotation – distances are unchanged by a unitary transformation. A unitary matrix $U$ obeys $UU^* = U^*U = I$. The singular value decomposition is widely used in numerical linear algebra as the “rotate – stretch – rotate” sequence avoids ill-conditioning. It is also useful for approximation techniques such as least squares, as small singular values can be set to zero to approximate a matrix by one of lower rank.

For transfer functions, the equivalent decomposition is

$$K(j\omega_1, j\omega_2) = U(j\omega_1)S V^*(j\omega_2),$$

or expressing this as a sum,

$$K(j\omega_1, j\omega_2) = \sum_i \sigma_i u_i(j\omega_1) v_i(j\omega_2).$$

The conditions for $U(j\omega)$ and $V(j\omega)$ to be unitary are chosen as

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} U(j\omega)U^*(j\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} U^*(j\omega)U(j\omega) d\omega = I \quad (2.14)$$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} V(j\omega)V^*(j\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} V^*(j\omega)V(j\omega) d\omega = I. \quad (2.15)$$
One should think of $\omega_1$ and $\omega_2$ as an infinite-dimensional analog of matrix subscripts. In general, $S$ is a diagonal operator, although for certain $K(j\omega_1, j\omega_2)$ it is finite-dimensional and hence a diagonal matrix.

**Singular value decomposition of the dynamic range limitation**

Suppose we apply the singular value decomposition to the dynamic range limitation multiplied by input weight $\Phi(j\omega)$ and output weight $\Psi(j\omega)$:

$$
\Psi(j\omega_1) \frac{H(j\omega_1) - H(j\omega_2)}{j(\omega_2 - \omega_1)} \Phi(j\omega_2) = \Psi(j\omega_1) G(j\omega_1) F(j\omega_2) \Phi(j\omega_2) = U(j\omega_1) S V^*(j\omega_2)
$$

(2.16)

Then we have the filter realisation

$$
F(j\omega) \Phi(j\omega) = V^*(j\omega)
$$

(2.17)

$$
\Psi(j\omega) G(j\omega) = U(j\omega) S.
$$

(2.18)

Why do this? Recall that $F(j\omega)$ is the gain to states and $\Phi(j\omega)$ can be thought of as the input signal to the filter. By setting $F(j\omega)\Phi(j\omega)$ equal to a unitary quantity $V^*(j\omega)$, we are guaranteeing that there will be unit average power in each state. This is a sensible scaling of signal level. There are further considerations to the effect that the gains to states and noise gains are uncorrelated, which are discussed in section 2.9.3. As the reader may have guessed from their name, the elements $\sigma_i$ of the diagonal matrix $S$ turn out to be the Hankel singular values $\sqrt{\lambda_i(K^\Phi W^\Psi)}$. Showing this is also deferred to section 2.9.3.

The values of the state-space matrices can be calculated. We use the property that

$$
\lim_{\omega \to \infty} (j\omega I - A)^{-1} = \frac{1}{j\omega I}
$$

and the definitions (2.3), (2.4) and (2.5) to calculate that

$$
D = \lim_{\omega \to \infty} H(j\omega)
$$

(2.19)

$$
C = \lim_{\omega \to \infty} j\omega G(j\omega)
$$

(2.20)

$$
B = \lim_{\omega \to \infty} j\omega F(j\omega).
$$

(2.21)
Calculating $A$ is a little less trivial. We use

$$F(j\omega) = (j\omega I - A)^{-1}B$$

$$(j\omega I - A)F(j\omega) = B$$

$$AF(j\omega) = j\omega F(j\omega) - B$$

$$AF(j\omega)\Phi(j\omega) = (j\omega F(j\omega) - B)\Phi(j\omega)$$

and then using (2.17),

$$AV^*(j\omega) = j\omega V^*(j\omega) - B\Phi(j\omega).$$

We post-multiply by $V(j\omega)$ and integrate over all frequency so that we can use (2.15):

$$A = \frac{1}{2\pi} \int_{-\infty}^{\infty} AV^*(j\omega)V(j\omega) \, d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} [j\omega V^*(j\omega) - B\Phi(j\omega)]V(j\omega) \, d\omega \quad (2.22)$$

The reader may have noticed that there is a big problem here, namely that the realisation $(A, B, C, D)$ is infinite-dimensional. However, it turns out that:

- If the transfer function being realised is rational and of order $n$, only $n$ singular values are non-zero.

- If only $n$ singular values are non-zero, $F(s)$ and $G(s)$ are those of an $n$th order state-space system.

- If the transfer function closely approximates that of an $n$th order system, then all but $n$ singular values will be small. If these small singular values are set to zero, the transfer function realised will be close to that desired. For weights $\Phi(s) = \Psi(s) = 1$, a rigorous error bound is available, namely that the maximum transfer function error is less than the sum of the singular values set to zero. For arbitrary weights, the property is only known qualitatively.

These properties are difficult to establish. Indeed, the author is not even sure that they have been proved for the frequency-weighted case. For the unweighted $(\Phi(s) = \Psi(s) = 1)$ case, they were proved by the Russian mathematicians Adamyan, Arov & Krein (1971). Good (but by no means simple) descriptions are (Glover 1984) and (Zhou 1996, p. 159–).
The author would be surprised if anything went wrong with the derivation used here, as it was found by transforming existing realisation theory into the frequency domain (see section 2.11) and including the frequency weights. Adding the weights should not cause problems as the resulting frequency-weighted realisation is also well known. As engineering verification, the author has tried the algorithm described below on many transfer functions (including high-pass and bandpass, and non-rational weights) and found no unexpected problems.

2.9.2 Numerical realisation algorithm

Quite remarkably, the equations above are actually a practical way of calculating a realisation. The transfer functions and operators have to be evaluated at a set of discrete frequency points to reduce the problem to a finite number of dimensions. For a filter with a cutoff frequency around 1 rad/s, a good choice might be

\[
\begin{align*}
\omega_1 &= -10 : 0.1 : 10 \\
\omega_2 &= -9.95 : 0.1 : 9.95.
\end{align*}
\]

(This notation means that \(\omega_1 = -10, -9.9, -9.8, \cdots, 9.9, 10\) etc.) We then calculate the \(201 \times 200\) matrix

\[
\mathcal{H} = \Psi(j\omega_1)\frac{H(j\omega_1) - H(j\omega_2)}{j(\omega_2 - \omega_1)}\Phi(j\omega_2).
\]

Note that we avoided \(\omega_2 = -10 : 0.1 : 10\) to eliminate problems at \(\omega_1 = \omega_2\). A alternative approach is to calculate \(\Psi(j\omega)\frac{dH(j\omega)}{dj\omega}\Phi(j\omega)\) for this diagonal.

Calculating the singular value decomposition of \(\mathcal{H}\) is a standard problem in numerical linear algebra (Golub & Van Loan 1996). In Matlab, the routines \textit{svd} or \textit{svds} can be used. The routine \textit{svds} produces only the \(n\) largest singular values and vectors, where \(n\) can be chosen. The resulting \(U\) and \(V\) are rectangular. If these are used, the realisation produced automatically only has order \(n\). (For the example which follows in section 2.9.4, \(U\) is \(200 \times 6\), \(V\) is \(201 \times 6\), and \(S\) is \(6 \times 6\).)

The matrices \(U\), \(S\) and \(V\) which result are not quite what we need. Firstly, \(U\) and \(V\) are only determined up to an arbitrary phase shift, which must be removed. This can be done
by calculating the unitary diagonal matrix

\[ M = \sqrt{U^T(j\omega)U(-j\omega)} \]

and substituting

\[
\begin{align*}
U &\leftarrow UM^{-1} \\
V &\leftarrow VM^*. \\
\end{align*}
\]

This substitution ensures that the realisation is real. Secondly, we need to normalise to satisfy (2.14) and (2.15) rather than \(UU^* = VV^* = I\). This is readily accomplished by

\[
\begin{align*}
U &\leftarrow \sqrt{\frac{2\pi}{\Delta\omega}} U \\
V &\leftarrow \sqrt{\frac{2\pi}{\Delta\omega}} V \\
S &\leftarrow \frac{\Delta\omega}{2\pi} S
\end{align*}
\]

where \(\Delta\omega\) is the frequency step – here \(\Delta\omega = 0.1\).

\(F(j\omega)\) and \(G(j\omega)\) can then be calculated using (2.17) and (2.18). To determine \(D, C\) and \(B\) using (2.19), (2.20) and (2.21), the author uses the approximations

\[
\begin{align*}
D &= \frac{1}{2} [H(-j10) + H(j10)] \\
C &= \frac{1}{2} [-j10 G(-j10) + j10 G(j10)] \\
B &= \frac{1}{2} [-j9.95 F(-j9.95) + j9.95 F(j9.95)]
\end{align*}
\]

There are presumably more accurate ways of doing this calculation. The matrix \(A\) is calculated using the Riemann sum to approximate the integral (2.22).

### 2.9.3 Properties of the optimal realisation

Using (2.17), (2.18), (2.14) and (2.15), and the definition of the Gramians (2.10) and (2.11),

\[
\begin{align*}
K^\Phi &= \frac{1}{2\pi} \int_{-\infty}^{\infty} [F(j\omega)\Phi(j\omega)][F(j\omega)\Phi(j\omega)]^* d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} V^*(j\omega) V(j\omega) d\omega \\
&= I
\end{align*}
\]
\[ W^\Psi = \frac{1}{2\pi} \int_{-\infty}^{\infty} [\Psi(j\omega)G(j\omega)]^* [\Psi(j\omega)G(j\omega)] \, d\omega \]
\[ = \frac{1}{2\pi} \int_{-\infty}^{\infty} S^* U^T(j\omega)U(j\omega)S \, d\omega \]
\[ = S^* S \begin{bmatrix} \sigma_1^2 & 0 & 0 & \ldots \\ 0 & \sigma_2^2 & 0 & \ldots \\ 0 & 0 & \sigma_3^2 & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{bmatrix} \]

The realisation will be called the frequency-weighted input-normal realisation. "Input normal" means \( K^\Phi = I \). In reading other literature, a little care should be taken as other input-normal realisations which do not have diagonal \( W^\Psi \) exist.

The frequency-weighted input-normal realisation has gains to states and noise gains which are uncorrelated under the frequency weights \( \Phi(j\omega) \) and \( \Psi(j\omega) \). As the eigenvalues of a diagonal matrix are the diagonal elements, the Hankel singular values \( \sqrt{\lambda_i(K^\Phi W^\Psi)} \) are clearly the \( \sigma_i \). The realisation is optimal in the sense that it minimises a measure of dynamic range \( \prod_i k_{ii}^\Phi w_{ii}^\Psi \). Recall that \( k_{ii}^\Phi \) is the frequency-weighted power gain to the \( i \)th state, and \( w_{ii}^\Psi \) is the frequency-weighted noise gain from the \( i \)th state to the output. For digital filters, this can be shown to minimise the total number of bits required to store the state to any given accuracy (Mullis & Roberts 1976). This fact will be used in an interesting calculation in chapter 5.

For analog filters, the correct measure we want to minimise is probably \( \sum_i w_{ii}^\Psi \) subject to \( k_{ii}^\Phi \leq 1 \) rather than \( \prod_i k_{ii}^\Phi w_{ii}^\Psi \). However, the frequency-weighted input-normal realisation is known to be fairly close to optimum in this measure also. It is possible to find a true minimum \( \sum_i k_{ii}^\Phi w_{ii}^\Psi \) by brute-force optimisation, but the realisation does not have any other conceptual significance.

There is a related realisation, called the frequency-weighted balanced realisation, which
has

\[ K^* = W = \begin{bmatrix}
\sigma_1 & 0 & 0 & \cdots \\
0 & \sigma_2 & 0 & \cdots \\
0 & 0 & \sigma_3 & \cdots \\
\cdots & \cdots & \cdots & \cdots
\end{bmatrix}. \]

It was originally found by Enns (1984) and Thiele (1986). It is trivial to obtain this from the input-normal realisation by diagonal scaling, so these authors were the first to derive the frequency-weighted input-normal realisation presented here. (They used a different algorithm – see section 2.11.4.) The frequency-weighted balanced realisation is useful for model reduction as states with small Hankel singular values \( \sigma_i \) can be deleted without affecting the transfer function much.

The dynamic range limitation can be used to derive a frequency-domain formula for the weighted output noise power \( \sum_i w_{ii}^{\Psi} \) achieved by the input-normal realisation. As \( K^* = I \),

\[ \sum_i w_{ii}^{\Psi} = \text{tr} K^* W^{\Psi}. \]  

(The trace \( \text{tr} \) is the sum of the diagonal elements, which also equals the sum of the eigenvalues.) From the definitions of the Gramians,

\[
\text{tr} K^* W^{\Psi} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[ \Psi(j\omega) G(j\omega) \right]^* \Psi(j\omega) G(j\omega) \, d\omega.
\]

Then using the commutivity of the trace, \( \text{tr} XY = \text{tr} YX \), with \( X = [F(j\omega_2) \Phi(j\omega_2)] \),

\[
\text{tr} K^* W^{\Psi} = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{tr} \left[ [F(j\omega_2) \Phi(j\omega_2)]^* \Psi(j\omega_1) G(j\omega_1) \right] [F(j\omega_2) \Phi(j\omega_2)] \, d\omega_1 d\omega_2.
\]

\[
= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{tr} \left[ [\Psi(j\omega_1) G(j\omega_1) F(j\omega_2) \Phi(j\omega_2)]^* [\Psi(j\omega_1) G(j\omega_1) F(j\omega_2) \Phi(j\omega_2)] \right] \, d\omega_1 d\omega_2.
\]

\[
= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{tr} \left[ [\Psi(j\omega_1) H(j\omega_1) - H(j\omega_2) \Phi(j\omega_2)]^* [\Psi(j\omega_1) H(j\omega_1) - H(j\omega_2) \Phi(j\omega_2)] \right] \, d\omega_1 d\omega_2.
\]

64
This is really just the dynamic range limitation as a weighted integral over input and output frequency. This is important, because it shows that the dynamic range as limited by the dynamic range limitation is in some sense achievable, namely by the frequency-weighted input-normal realisation. For readers unfamiliar with matrix transfer functions, the simplification for scalar (single-input, single-output) transfer functions may be easier to comprehend:

\[
\operatorname{tr} K^\Phi W^\Psi = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\Phi(j\omega_2)|^2 |\Psi(j\omega_1)|^2 \left| \frac{H(j\omega_1) - H(j\omega_2)}{j(\omega_2 - \omega_1)} \right|^2 d\omega_1 d\omega_2
\]

### 2.9.4 Example of frequency-weighted input-normal realisation

As an example, we realise the same 1 rad/s low-pass 5th-order Butterworth transfer function as in section 2.5.1. We use weightings

\[
\Phi(s) = \frac{(1 + s)^2}{(1 + 0.25s)^2} \\
\Psi(s) = 1.
\]

These are quite realistic for a radio receiver. The input weight \(\Phi(s)\) has two real zeros at 1 rad/s and two real poles at 4 rad/s. This represents an input signal which rises from 0 dB in-band to +24 dB out-of-band. Output noise is considered equally important at all frequencies.

As the transfer function is fifth-order, only five Hankel singular values should be non-zero. However, we decide to keep six to investigate how accurate the algorithm is. The magnitudes of the transfer function \(H(j\omega)\), gains to states \(F(j\omega)\) and noise gains \(G(j\omega)\) are shown in fig. 2.5. A wide frequency and amplitude range are shown, so as to show imperfections.

The high-frequency error in the realised transfer function is due to the frequency grid only extending up to 10 rad/s, leading to inaccurate values of \(B, C\) and \(D\). Interestingly, the number of points does not need to be large for an accurate transfer function – a 10 \times 9 matrix gives the same transfer function as a 201 \times 200 matrix. (The gains to states and noise gains are badly suboptimal though.)

The \(F(j\omega)\) for states 1–5 can be seen to be sensibly scaled. The increased input
Figure 2.5: Transfer function, gains to states and noise gains of the “optimal” realisation of a fifth-order Butterworth response. The sixth state (black) is only included for illustrative purposes.
weighting above 1 rad/s lowers the gains to states there, although the falloff in the transfer function also contributes. The magnitudes of the noise gains $G(j\omega)$ decrease in accordance with the decreasing Hankel singular values as we move from $g_1$ to $g_5$. The first state has a higher noise gain than in the LC ladder realisation. This is a consequence of the gains to states being lower out-of-band than those of the LC ladder realisation. (Compare fig. 2.5 and fig. 2.4.)

The behaviour of the sixth state is interesting. We would expect the noise gain $g_6$ to be zero as the sixth frequency-weighted Hankel singular value should be zero, and indeed it is quite small. It is not zero because of rounding and numerical routine inaccuracies, and because of the two approximations in the algorithm – the finite number of frequency points, and the calculation of $B$ (and indirectly, $A$), $C$ and $D$ from values at $10$ rad/s rather than $\infty$. The error varies considerably from run to run, which suggests that the Matlab singular value decomposition algorithm $svds$ uses iteration with a random initial guess.

Approximating a rational transfer function by a higher-order rational transfer function nearly always produces almost coincident pole-zero pairs, and this algorithm is no exception. Here the pole-zero pair is very close to the origin, resulting in the low-frequency peak in $f_6$.

In this and other examples, the algorithm worked reliably when used with a sensible matrix size and frequency range. With $201 \times 200$ matrices the calculations takes a few seconds. It should be noted that existing matrix-based approaches (Zhou 1996, p. 163) (Anderson & Moore 1989, problem 10.2-1) (Enns 1984) (Thiele 1986) are faster and more accurate if state-space realisations of the transfer function and weights are available. This other approach is briefly introduced in section 2.11.4.

2.10 TRANSFER FUNCTION OPTIMISATION

What transfer function optimises the dynamic range, subject to the dynamic range limitation? In this thesis, we consider this for two cases, filters and amplifiers. Filters are deferred to the next chapter. The brief summary for filters is that (2.9) seems to summarise the situation – the achievable dynamic range is limited by how close the stopband is to the passband and little else. There is surprisingly little difference between, say, an Elliptic and
a Chebyshev response. (Sensitivity depends a lot more on transfer function than dynamic range though, as will be shown in the next chapter.) Amplifiers are discussed in this section.

2.10.1 Amplifier transfer function

If a transfer function is constant, (2.8) is zero and the dynamic range is unlimited. For a filter, obviously \( H(s) = c \) is useless. For an amplifier, \( H(s) = c \) with \(|c| > 1\) would be nice, but it is not possible because of the limited speed of the transistors. At frequencies above \( f_{max} \), the power gain must be less than 0 dB (Thornton, DeWitt, Chenette & Gray 1966, p. 114). We will investigate the example of an amplifier with a low-frequency power gain of 20 dB. The transfer function must therefore vary with frequency, and hence the dynamic range limitation (2.8) must be non-zero.

In most applications, maximum signal and noise and distortion matter more at some frequencies than others. The weights \( \Phi(s) \) and \( \Psi(s) \) and the Gramians were introduced to deal with this problem, but for simplicity we work directly from the dynamic range limitation in this section. We will assume that the aim is to maximise dynamic range over a band starting at DC and extending up to a frequency \( \omega_P \). Transistor \( f_{max} \) and other considerations (\( f_T \), load capacitance etc.) require that the gain of the amplifier fall to 0 dB by \( \omega_C \).

According to (2.8), we should minimise the magnitude of \( [H(j\omega_1) - H(j\omega_2)]/j(\omega_2 - \omega_1) \) over DC to \( \omega_P \). A straightforward way of doing this is to require \( H(j\omega) \) to be close to constant over that frequency range. Note that \( H(j\omega) \) needs to be constant in phase as well as magnitude – this is not the standard filter approximation problem. The target was chosen as \( H(j\omega) = 10 + j0 \). (The imaginary part has to be zero to be realisable at DC.)

The optimisation as described above tends to produce an \( H(j\omega) \) which has a large peak between \( \omega_P \) and \( \omega_C \). This is likely to be intolerable in a practical amplifier, so a constraint on maximum \( |H(j\omega)| \) needs to be imposed. \( |H(j\omega)| < 14 \) was used here, limiting peaking to 3 dB. (It may be desirable to limit the magnitude of \( [H(j\omega_1) - H(j\omega_2)]/j(\omega_2 - \omega_1) \) as well. However, the peaking constraint alone gave reasonable results.)
The transfer function $H(j\omega)$ also needs to be stable and causal. This is obvious but mathematically important.

To summarise, the problem is:

Minimise (the maximum value of) $|H(j\omega) - 10|$, $0 \leq \omega \leq \omega_P$

for $H(j\omega)$ stable and causal and subject to $|H(j\omega)| < 14$ for all $\omega$ and $|H(j\omega)| < 1$ for $\omega > \omega_C$.

There are a variety of ways to solve this optimisation problem. It can be formulated as trying to find a causal approximation to the “centre”

$$C(j\omega) = \begin{cases} 
10, & |\omega| < \omega_P \\
0, & |\omega| \geq \omega_P 
\end{cases}$$

within an error bound radius

$$R(j\omega) = \begin{cases} 
k, & |\omega| < \omega_P \\
14, & \omega_P < |\omega| < \omega_C \\
1, & |\omega| \geq \omega_C 
\end{cases}$$

where we wish to minimise $k$. $C(j\omega)$ is clearly non-causal as a filter with the same magnitude response would have an associated phase lag in accordance with the Bode (1945) gain-phase relationships. $C(j\omega)$ as it stands has zero phase. In this formulation the problem basically comes down to making a weighted (by $R(j\omega)$) causal approximation to $C(j\omega)$. It is discussed at length by Helton (1981). (The problem of making such causal approximations is called the Nehari problem by mathematicians.) Fortunately, Helton & Merino (1998) have written Mathematica procedures to solve the problem.

There are other solutions. Brute-force approximation, say over a space of FIR transfer functions, can be quite practical. (The discrete-time FIR responses can be converted to continuous-time by a bilinear transformation.) As we shall see, optimising the loop gain of a feedback amplifier also basically solves the same problem. However, Merino and Heltons’ Mathematica routines seems to be faster and more consistent, so we use them. They also give an indication of how close to optimum the approximation achieved is. The output of
these routines is the optimum frequency response $H(j\omega)$ over a few hundred frequency points.

Typical results $H(j\omega)$ of the optimisation are shown on the left in fig. 2.6. These (the blue, green and red curves) are for $\omega_P = 1$ rad/s and $\omega_C = 5, 10, 20$ rad/s. It may occur to the reader that the curves look like the closed-loop frequency responses of feedback amplifiers. (This resemblance is pretty feeble at present; they looked more like feedback amplifiers in earlier optimisation runs before Helton’s superior algorithm was used.) They are very flat and have little phase shift in-band (as $H(j\omega) \simeq 10 + j0$ for $\omega < \omega_P$), and have 3 dB of (peculiar-shaped) peaking out-of-band (as constrained by $|H(j\omega)| < 14$). This makes sense, for we suspect that feedback amplifiers have high dynamic range. The signal transfer function of a high-order single-loop sigma-delta modulator also has a similar shape.

If the responses were those of feedback amplifiers, it is reasonable to ask what the open-loop gain would be. This can be calculated from

$$H(s) = \frac{L(s)}{1 + kL(s)} \Rightarrow L(s) = \frac{H(s)}{1 - kH(s)}$$

where $L(s)$ is the opamp gain and $k$ is the feedback factor. (“Opamp gain” just means open-loop forward gain here.) As the amplifier has 20 dB gain, we use $k = 0.1$. This results in the opamp gains $L(s)$ on the right of fig. 2.6. It is certainly conceivable that an “optimum” feedback amplifier could have such an opamp gain. It may be easier to think of rad/s as GHz – then these are opamps compensated for a closed-loop gain of 10 over 1 GHz bandwidth, where the opamps have 5, 10, 20 GHz unity-gain bandwidth.

The phase margins are around 50° – remember that the loop gain falls to one where the opamp gain is 20 dB as $k = 0.1$. This phase margin is primarily due to limiting peaking to 3 dB. More peaking allows more loop gain below $\omega_P$ at the expense of reduced phase margin. For $\omega_C = 10$ rad/s, doubling the peaking to 6 dB allows 77 dB opamp gain below $\omega_P = 1$ rad/s, as compared to 60 dB for the green curve. The flat low frequency gain is obviously the result of optimisation. The higher it is, the closer $H(j\omega)$ is to $10 + j0$.

Gratuitous peaks in the low-frequency loop gain tend to get smoothed out as they degrade the phase margin in accordance with the various Bode gain-phase relations (Bode 1945). The responses are conditionally stable, albeit only just for the blue curve. This is normally
Figure 2.6: Optimised transfer functions $H(s)$ and corresponding open-loop gains $L(s)$. Blue, green, red: $\omega_C = 5\, \text{rad/s}, 10\, \text{rad/s}, 20\, \text{rad/s}$. Cyan: one-pole rolloff above $\omega_C = 10\, \text{rad/s}$. 
avoided in amplifiers, but it is known to allow greater low-frequency loop gain (Oizumi & Kimura 1957) and is used in high-order single-loop sigma-delta modulators.

The assumption that an amplifier can have a gain very close to unity at all frequencies above $\omega_S$ may be suspect. For this reason, the cyan curve in fig. 2.6 is included. In this case, $|H(j\omega)|$ was constrained to fall off at a one-pole rate at high frequency. It can be seen that this reduces the in-band opamp gain $L(j\omega)$ from 60 dB (the green curve) to 49 dB (the cyan curve). Constraining gain to fall off at a two-pole rate causes a further deterioration to around 40 dB low-frequency opamp gain, i.e. 20 dB of feedback. It may be of some significance that the very fast opamps in the next chapter have a gain which falls off at only a one-pole rate at high frequencies, due to a feedforward stage. Perhaps other “feedforward” techniques such as gain-boosted cascodes also have wide bandwidth because they reduce the order of the high-frequency asymptotic gain rolloff?

### 2.10.2 What is feedback?

The optimisations in the section above and the feedback amplifier example in section 2.5.2 provide a distinctive view of feedback circuits. It seems that a feedback circuit “works” by minimising the dynamic range required of its active components within a passband (the frequencies where loop gain is high). The following comment from Bode (1945, p. 47) indicates that this point has been known for a long time:

“... the usual conception of feedback includes two distinct ideas. The first is that of a loop transmission or return of voltage, and the second that of a reduction in the effects of variation in the tube characteristics. In normal circuits these two are related by simple mathematical laws so that the term “feedback” can refer generically to both.

In exceptional circuits, when the correlation between the two breaks down, the first idea is evidently the one which most nearly agrees with the usual physical conception of feedback. It will therefore be taken as the basis for feedback in the general case.”
But what if we abandon the first idea, and work entirely from the second? The conventional definition of loop gain using return difference and signal flow graphs (Bode 1945) (Mason 1954) is somewhat cumbersome. It would be better if the definition could be related back to nodal analysis or some other unambiguous circuit description. If active circuit synthesis was sufficiently developed, could we synthesise “feedback” circuits by minimising $F(j\omega)G(j\omega)$ over the desired frequency range rather than by creating a feedback path?

**Compensation**

The dynamic range limitation (2.8) shows that in a circuit with restricted bandwidth, at least one state must be stored with decent dynamic range. Looking back to the feedback amplifier example (section 2.5.2), we see that the state $x_4$, the voltage across the compensation capacitor, serves this function. The other three states all have reduced dynamic range requirements. Thus, instead of using a wide bandwidth and optimising the dynamic range (as in the optimisation example, section 2.10), we also have the option of making a distinction between states associated with transistors and hence distortion, and states associated with passive components. The dynamic range only needs to be minimised for the “transistor” states.

As compensation using linear capacitors is widely used, it would be useful to have an optimisation method which could handle two classes of states – low dynamic range “transistor” states and high dynamic range “capacitor” states. However, the author does not have a method for doing this at present. If one is developed, transistor-level active circuit synthesis becomes a real possibility.

### 2.11 LINKS TO OTHER LITERATURE

This chapter takes a distinctive frequency-domain approach to the state-space dynamic range theory. The key is to use two frequency variables $\omega_1$ and $\omega_2$ and the dynamic range limitation. The first priority of this chapter is to present this theory without diversion, as a combination of new concepts and new mathematics is usually difficult to understand.
However, the approach based on the dynamic range limitation did not come from nowhere - to do this would be a task of similar magnitude to the original development of state-space theory by Kalman, Bellman and others in the 1960’s. Essentially, the author came up with the dynamic range limitation, and some months later derived that it was the Laplace transform of the so-called Hankel operator of time-domain realisation theory. This existing theory was then used as a guide to developing the realisation in section 2.9. This section introduces the existing theory so that new work can be distinguished from old.

2.11.1 The Hankel operator

The dynamic range limitation is the frequency-domain representation of the Hankel operator. This does not seem to be existing knowledge. We need to start by mentioning what an operator is.

We can think of an input or output signal as an infinite vector consisting of the values at all the different times or frequencies. An operator is an infinite matrix which maps the input to the output. (By operator we mean linear operator.) Mathematicians take a dislike to the term “infinite matrix” as the key to operator theory is to separate the operators from their matrix representation in a particular basis, and to consider finite-dimensional and infinite-dimensional quantities in similar ways. A function is like a diagonal operator. A time-domain convolution is a good example of a non-diagonal operator. A Fourier or Laplace transform is also an operator. Note the analogy between a discrete Fourier transform (DFT), which is a finite matrix, and the continuous Fourier transform which is an infinite matrix or operator. As an aside, note that the Laplace or Fourier transform diagonalises the convolution operator, giving the transfer function. This is why the frequency domain is so useful for analysing linear time-invariant systems.

The Hankel operator $\mathcal{H}$ is the operator which maps past inputs to future outputs. A standard reference on it is (Glover 1984). For a time-invariant system with impulse response $h(t)$,

$$ (\mathcal{H}u)(t) = \int_{-\infty}^{0} h(t - \tau)u(\tau)d\tau, \quad t > 0 $$
It is convenient to define the input to the Hankel operator as \( v(t) = u(-t), \ t > 0 \) rather than \( u(t), \ t < 0 \) (Glover 1984). This is useful here as we are going to Laplace transform the above expression, and the Laplace transform \( \mathcal{L} \) is defined as

\[
V(s) = (\mathcal{L}v)(s) = \int_{0}^{\infty} e^{-st}v(t)dt
\]

which ignores \( t < 0 \). Thus we use

\[
(\mathcal{H}v)(t) = \int_{0}^{\infty} h(t + \tau)v(\tau)d\tau, \ t > 0.
\]

(The Hankel operator is still only defined for output time \( t > 0 \); it is the input we have time-reversed, not the output.)

We will need expressions for the impulse responses of the transfer function, gain to states, and noise gain. These are readily derived from the state-space equations or from the inverse Laplace transforms of \( H(s), F(s) \) and \( G(s) \) to be

\[
\begin{align*}
h(t) &= Ce^{At}B + D\delta(t) \\
f(t) &= e^{At}B \\
g(t) &= Ce^{At}.
\end{align*}
\]

For simplicity we have assumed zero initial conditions. \( \delta(t) \) is a unit impulse at time \( t = 0 \). We assume \( D = 0 \) for now, although it seems possible to neglect this. Therefore

\[
(\mathcal{H}v)(t) = \int_{0}^{\infty} Ce^{A(t+\tau)}Bv(\tau)d\tau
\]

\[
= \int_{0}^{\infty} Ce^{At}e^{A\tau}Bv(\tau)d\tau
\]

\[
= Ce^{At} \int_{0}^{\infty} e^{A\tau}Bv(\tau)d\tau
\]

\[
= g(t) \int_{0}^{\infty} f(\tau)v(\tau)d\tau.
\]

This can be Laplace transformed to give

\[
(\mathcal{H}V)(j\omega) = \frac{1}{2\pi} G(j\omega_1) \int_{-\infty}^{\infty} F(j\omega_2)V(-j\omega_2)d\omega_2
\]

75
using the generalised Parseval’s formula on the integral, and with $\omega_1$ the frequency variable for $t$, and $\omega_2$ the frequency variable for $\tau$. (Doetsch (1974, p. 211) gives a good discussion of Parseval’s formula for Laplace transforms.) Then using the dynamic range limitation (2.8),

$$ (HV)(j\omega_1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{H(j\omega_1) - H(j\omega_2)}{j(\omega_2 - \omega_1)} V(-j\omega_2) d\omega_2. \quad (2.24) $$

We see that the dynamic range limitation is the kernel of the Laplace-transformed Hankel operator. This is why we were justified in calling the singular values of the dynamic range limitation Hankel singular values.

### 2.11.2 Defining a frequency-weighted Hankel operator

In light of (2.24), it seems natural to define a frequency-weighted Hankel operator

$$ (HV)(j\omega_1) = \frac{1}{2\pi} \Psi(j\omega_1)G(j\omega_1) \int_{-\infty}^{\infty} F(j\omega_2)\Phi(j\omega_2)V(-j\omega_2) d\omega_2 \quad (2.25) $$

$$ = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi(j\omega_1)H(j\omega_1) - H(j\omega_2)\Phi(j\omega_2)V(-j\omega_2) d\omega_2. \quad (2.26) $$

The next section shows that this quantity has the same properties as the unweighted version.

### 2.11.3 Controllability and observability operators

Recall that the Hankel operator $\mathcal{H}$ maps past inputs to future outputs. The reason that it is a useful quantity is that it can be factorised into a map from past input to current state and a map from current state to future output (Glover 1984) (Zhou 1996). This step is important for state-space realisation. These maps are known as the controllability operator $C$ and the observability operator $O$ respectively. They are known to be given by (Zhou 1996)

$$ C(v(t)) = \int_{0}^{\infty} e^{A\tau}Bv(\tau) d\tau $$

$$ = \int_{0}^{\infty} f(\tau)v(\tau) d\tau $$

$$ (Ox)(t) = Ce^{At}x $$

$$ = g(t)x $$
It is straightforward to Laplace transform these expressions, but the author has never seen it done. Presumably, without the dynamic range limitation there is little reason to do so. The result is

\[
C(V(j\omega)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega)V(-j\omega)d\omega
\]

\[
(\mathcal{O}X)(j\omega) = G(j\omega)X.
\]

In light of this, it is natural to define frequency-weighted controllability and observability operators

\[
\mathcal{C}(V(j\omega)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega)\Phi(j\omega)V(-j\omega)d\omega
\]

\[
(\mathcal{O}X)(j\omega) = \Psi(j\omega)G(j\omega)X.
\]

These expressions satisfy the same properties as the unweighted time-domain expressions, which are (Zhou 1996)

\[
\mathcal{H} = \mathcal{O}C
\]

\[
K^\Phi = CC^s
\]

\[
W^\Psi = \mathcal{O}^s\mathcal{O}.
\]

### 2.11.4 Optimum dynamic range realisation

Section 2.9 described a new approach for calculating “optimal” realisations. Here we briefly describe what existed before. There were two main approaches. One was based on similarity transformation of an existing state-space realisation. This could deal with frequency weightings. The other was closer to that in section 2.9, but used the singular value decomposition of the Hankel operator in the time domain. This makes it more difficult to apply frequency weighting, and the author has not seen a frequency-weighted time-domain algorithm.

**Similarity transform approach**

We saw in section 2.8.2 that the different minimal state-space realisations of a transfer function are parameterised by a similarity transformation \(T\). The traditional approach to
optimum dynamic range realisation is to take any realisation, good or bad, and find a $T$
which transforms it into an optimal one.

For any realisation, there exists a $T$ which transforms it into one with the Gramians $K^\Phi$
and $W^\Psi$ diagonal. (The $T$ of course depends on the realisation you start with.) The
existence of such a similarity transformation which simultaneously diagonalises two
positive-real symmetric matrices is a moderately well-known piece of linear algebra.

The desired transform $T$ can be produced using a Cholesky decomposition
(upper-triangular $R$)

$$ R^* R = W^\Psi, $$

(2.27)
a Schur decomposition (unitary $U$, upper-triangular $S$)

$$ U S U^* = R K^\Phi R^*, $$

(2.28)
and two square roots and an inverse,

$$ T = S^{-1/4} U^* R. $$

(2.29)

This is taken from (Anderson & Moore 1989, problem 10.2-1). Thiele (1986) gives a similar
procedure. This $T$ actually results in the frequency-weighted balanced realisation rather
than the frequency-weighted input-normal realisation, but this is easily corrected by
diagonally scaling the signal levels.

**Time-domain Hankel singular value decomposition approach**

That a state-space realisation could be derived by factoring the Hankel operator into
controllability and observability operators has been known for a long time. (In all practical
algorithms operators are represented by finite matrices.) The original algorithm is by Ho &
Kalman (1966). Ho and Kalman used a different matrix factorisation. The singular value
decomposition was later found to be convenient by Zeiger & McEwen (1974) and Kung
(1978).

Once the Hankel operator is factorised as $\mathcal{H} = OC$, the matrices $A, B, C$ and $D$ are
uniquely determined and can be calculated. The time-domain procedures for doing this are
of a similar nature to the frequency-domain formulae in section 2.9.1, but involve matrix shifts and the like.

In the time domain, the Hankel operator does not have an obvious frequency-weighted generalisation. Thus previous authors do not seem to have been extended the Ho-Kalman / singular value decomposition approach to the frequency-weighted case. This is the value of the frequency-domain approach presented in this chapter.

2.11.5 Sensitivity

Sensitivity is how much the transfer function varies when a component value varies.
Dynamic range is closely linked to sensitivity. This is not surprising, because a component with a changed value can always be represented as the original component plus a noise.

Deriving the change in transfer function is quite straightforward. For changes in $B$, $C$ and $D$,

$$H_{\Delta B}(s) - H(s) = [C(sI - A)^{-1}(B + \Delta B) + D] - [C(sI - A)^{-1}B + D]$$

$$= C(sI - A)^{-1}\Delta B$$

$$= G(s)\Delta B$$

$$H_{\Delta C}(s) - H(s) = [(C + \Delta C)(sI - A)^{-1}B + B] - [C(sI - A)^{-1}B + D]$$

$$= \Delta C(sI - A)^{-1}B$$

$$= \Delta C F(s)$$

$$H_{\Delta D}(s) - H(s) = [C(sI - A)^{-1}B + D + \Delta D] - [C(sI - A)^{-1}B + D]$$

$$= \Delta D.$$ 

For changes in $A$, we use a derivation remarkably like that for the dynamic range limitation:

$$M = \left(sI - (A + \Delta A)\right)^{-1} - (sI - A)^{-1}$$

$$(sI - (A + \Delta A))M(sI - A) = (sI - A) - (sI - (A + \Delta A))$$

$$= \Delta A$$

$$M = (sI - (A + \Delta A))^{-1}\Delta A(sI - A)^{-1}$$

79
Then

\[ CMB = \left[ C(sI - (A + \Delta A))^{-1}B + D \right] - \left[ C(sI - A)^{-1}B + D \right] \]

\[ = H_{\Delta A}(s) - H(s) \]

but also

\[ CMB = C(sI - (A + \Delta A))^{-1} \Delta A(sI - A)^{-1}B \]

\[ = G_{\Delta A}(s) \Delta A F(s). \]

Combining all these results for infinitesimal changes,

\[ dH(s) = G(s) dA F(s) + G(s) dB + dC F(s) + dD. \]

This result is well known in a less elegant subscripted form. Other authors (Thiele 1986) (Lutz & Hakimi 1988) (Snelgrove & Sedra 1986) calculate derivatives instead of using a resolvent identity.

Note that whereas the dynamic range limitation uses the gain to states and noise gain at two different frequencies, the sensitivity formulae use them only at one frequency. This is because a parameter change produces an output “noise” at the same frequency as the input. For this reason, there is more opportunity to optimise sensitivity than dynamic range by appropriate choice of the transfer function. This consideration will be discussed further in the next chapter.

2.11.6 Sensitivity theory and other related previous work

To the best of the author’s knowledge, the dynamic range limitation (2.8) is new. The closest result the author has seen is

\[ G(s)F(s) = \frac{dH(s)}{ds}, \]

derived by Snelgrove & Sedra (1986). This is a corollary of the dynamic range limitation for \( s_1 \to s_2 \). Note that this is the case which is relevant for sensitivity analysis. Historically, sensitivity has been a more important problem than dynamic range. There was great interest in the sensitivity problem in the late 1960’s and early 1970’s, and it would not be
surprising if the more general dynamic range limitation was known in this context. The author has read through the IEEE Transactions on Circuits and Systems for this period and cannot find the dynamic range limitation (even in sensitivity rather than state-space notation).

The dynamic range limitation may not have been discovered by sensitivity theorists because the resolvent identity was not widely known. However, the resolvent identity is widely used in functional analysis (Dunford & Schwartz 1964, p. 568). Zames (1981) mentions the use of resolvent identities for sensitivity analysis. It would not be at all surprising if Zames had derived the dynamic range limitation too, considering his depth of knowledge of functional analysis and control system robustness.

Livsic (1973, p. 29) uses a similar identity to derive some properties of lossless circuits, and to develop a variant of Darlington synthesis. Livsic’s $S(\omega)$ and $R(\omega)$ seem to be the same as $H(j\omega)$ and $F(j\omega)$ here. Livsic does not use a noise gain $G(j\omega)$.

2.11.7 Energy storage, sensitivity and group delay

Some understanding of the relationships between transfer function and dynamic range can be acquired by other means. In particular, the expressions which relate the stored energy in a lossless network to the group delay (Kishi & Nakazawa 1963) (Smith 1965) constrain $F(j\omega)$ (Harrison & Weste 2001). Sensitivity has also been previously related to group delay and stored energy (Kishi & Kida 1967) (Orchard et al. 1985) (Temes & Orchard 1973). It is also possible to relate $G(j\omega)$ to $F(j\omega)$ using reciprocity (Harrison & Weste 2001). However, the work presented here seems to give a more useful limitation, and is not limited to lossless reciprocal systems.

2.12 CONCLUSION

This chapter has presented a means of calculating the dynamic range of an active filter. The main benefit of the theory presented here is that it allows dynamic range and sensitivity to be taken into account when choosing the filter transfer function. Of course,
previous approaches such as checking the maximum pole $Q$ provide some insight, but the
dynamic range limitation and Gramians offer a rigorous approach which can cope with
transfer functions outside a designer’s experience.

There is a natural state-space structure implicit once the transfer function is chosen. Its
dimensions are found from the singular values of the Hankel operator. In the
frequency-domain, the Hankel operator has kernel $[H(j\omega_1) - H(j\omega_2)]/j(\omega_2 - \omega_1)$, which is
also equal to the product of the gain to states $F(j\omega_2)$ and the noise gain from the states
$G(j\omega_1)$. The frequency-weighted input-normal form uses this natural structure to achieve
high dynamic range.

The state-space dynamic range theory also provides some insight into feedback amplifiers.
Feedback can be thought of as a combination of a flat transfer function which allows high
dynamic range according to the dynamic range limitation, and a good realisation of this
transfer function. Opamps do not have quite as neat a state-space description as filters, so a
little more work is needed before they can be synthesised to optimise dynamic range.
To further investigate dynamic range considerations for deep-submicron CMOS circuits, an active filter was designed in a 0.18\textmu m process. The transfer function is a fifth-order Elliptic low-pass which would be useful as an anti-alias filter (such as in a broadband wireless LAN's baseband circuitry). The choice of the transfer function is a further example of the use of the dynamic range limitation, and one of real practical merit.

It would be conceptually elegant to use the frequency-weighted input-normal realisation discussed in the previous chapter. However, a more conventional realisation based on simulating an LC ladder was used. This has advantages with respect to sparsity of the tap matrix, as will be discussed in section 3.3.

The filter uses opamps as the active element. This is a somewhat radical choice for cutoff frequencies above 100 MHz; conventional wisdom would be to use a gm-C approach. An opamp-RC approach was tried as feedback circuits seemed promising for achieving large signal swings with low distortion. The theory in section 2.10 also suggested that feedback amplifiers had not been pushed to their bandwidth limits.

Resistors are switched in or out to tune the filter. (Unlike switched-capacitor circuits, continuous-time filters need to tune out the effects of resistor and capacitor absolute tolerance.) The general reason that resistors were preferred to a MOSFET-C approach is the better linearity with high signal swings. In systems on a chip, having switched rather than continuous tuning is not a problem as the tuning loop is usually digital and can easily deal with switched elements.
3.1.1 Organisation of the chapter

The conventional design process for a filter involves three key choices – choice of transfer function, choice of realisation, and choice of integrator (or resonator or biquad). Section 3.2 discusses choosing the transfer function. The dynamic range limitation is used. The section summarises and uses some of the theory in the previous chapter, and represents a good chance for the reader to further understand the use of that material. Section 3.3 discusses the realisation. (Note that in filter design, “realisation” basically means the state-space equations, not the transistor-level design.)

Section 3.4 describes the opamp. An opamp-based approach was chosen a-priori on the grounds that gm-C filters have already received plenty of attention by many distinguished authors. A comparison of the relative merits of the gm-C and opamp-RC approaches is deferred to the results section (3.5). The opamp work is the major new contribution in this chapter. The basic problem is to achieve sufficient loop gain at high frequencies.

After the measured results on the fabricated 350MHz filter, the chapter finishes with some further simulations to see if even greater bandwidth and signal swing is possible. It is, and as the fabricated filter is already state-of-the-art, the opamp-RC approach looks promising indeed.

3.2 CHOICE OF TRANSFER FUNCTION

Choosing the filter transfer function is a trade-off between the passband and stopband specifications on one hand, and the dynamic range and sensitivity on the other. A basic outcome of the previous chapter is that the transfer function \( H(j\omega) \) should be chosen to minimise \( |H(j\omega_1) - H(j\omega_2)|/j(\omega_2 - \omega_1) \). Some input signal frequencies \( \omega_2 \) and output noise frequencies \( \omega_1 \) are more important than others, namely the \( \omega_2 \) where the input signal is large or the \( \omega_1 \) where the output noise needs to be small. In section 2.11.5, sensitivity was shown to depend on values along the line \( \omega_1 = \omega_2 \), ie. on \( |dH(j\omega)/d(j\omega)| \).
3.2.1 Transfer function optimisation

A logical approach would be to find the transfer function $H(j\omega)$ which minimises

$$\text{tr} K \Phi W \Psi = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\Phi(j\omega_1)|^2 |\Psi(j\omega_2)|^2 \left| \frac{H(j\omega_1) - H(j\omega_2)}{j(\omega_2 - \omega_1)} \right| d\omega_1 d\omega_2$$

in accordance with the theory in section 2.9.3. The transfer function must of course be constrained to meet the passband and stopband specifications and to be causal and stable. Recall that $|\Phi(j\omega)|^2$ is the input power spectrum, and $|\Psi(j\omega)|^2$ is the weight to be put on the output noise. For an anti-alias filter in a wireless LAN receiver, $|\Phi(j\omega)|^2$ might be chosen as 0 dB in-band and +10 dB out-of-band to cope with 10 dB larger blocking signals.

The noise weight $|\Psi(j\omega)|^2$ might be chosen as 0 dB in-band and for frequencies which the ADC aliases down into the passband, and -10 dB for other frequencies. (It would probably be unwise to go below -10 dB, as the weights affect sensitivity versus frequency as well as dynamic range.)

The most straightforward way to minimise $\text{tr} K \Phi W \Psi$ would be to numerically optimise some parameterisation of $H(s)$. This was tried for rational $H(s)$ and for “FIR” $H(s)$ with $(1 - s)/(1 + s)$ in place of $z^{-1}$ as the delay element. The results were mixed. It seems fairly clear that the optimum is a transfer function with a rounded passband corner and not more stopband attenuation than is necessary. The results often look fairly similar to an inverse-Chebyshev response or an Elliptic with low passband ripple, sometimes with the stopband zeroes moved a little into the left half-plane. It is not possible to say much more than this, because the optimisation tends to have local minima and reflects the parameterisation chosen.

Values of $\text{tr} K \Phi W \Psi$ within 0.5 dB of the optimised results were possible by just choosing an Elliptic response with 0.01–0.1 dB passband ripple, and stopband attenuation and order to just meet the specification. Such a response, namely a 5th order Elliptic with 0.01 dB passband ripple and -49 dB stopband, was chosen for the filter described in this chapter.

The optimal response shape is largely independent of the weightings $|\Phi(j\omega)|^2$ and $|\Psi(j\omega)|^2$.

Realising stopband zeroes can be cumbersome, so it was checked whether an all-pole
response such as a Chebyshev transfer function could achieve similar performance. For weightings as above, the general result was that the best Chebyshev transfer functions had \( \text{tr} K^\Phi W^\Psi \) about 1.5 – 2 times that of the low-ripple Elliptic responses. This would result in 2–3 dB loss in dynamic range. The best Chebyshev responses had moderate ripple, 0.01–0.3 dB.

3.2.2 Graphical examination of dynamic range limitations

To compare different transfer functions, it can be useful to plot the quantity in the dynamic range limitation (the Hankel operator kernel) \( ||H(j\omega_1) - H(j\omega_2)|| / j(\omega_2 - \omega_1) \). Fig. 3.1 shows this for an Elliptic and a Chebyshev transfer function with the same transition bandwidth. The Elliptic response is that used in the filter described in this chapter. Both filters are normalised to -3 dB at 1 rad/s and are -49 dB at 1.63 rad/s.

Recall that sensitivity is related to performance along the line \( \omega_1 = \omega_2 \), as a parameter variation generates a "noise" at the input frequency. Blocking performance is for passband \( \omega_2 \), and stopband \( \omega_1 \). These regions are shown on fig. 3.1. The two frequency responses have almost identical dynamic range limits for a stopband blocking signal. This is a consequence of the results in section 2.6.3. However, the Elliptic response has about 7 dB lower passband sensitivity (no dark red region on the plots). (All of these results of course assume a good realisation.)

It is worth spending time studying fig. 3.1 in detail, as the plots bring together a lot of the theory in the previous chapter in a way which is likely to be useful to a designer.

3.2.3 Relationship with existing theory

It is useful to relate to this new approach to choosing the transfer function using the dynamic range limitation to existing work. It is known that minimising filter pole Q tends to minimise sensitivity and maximise dynamic range (Abidi 1992) (Groenewold 1991). In particular, the high-Q pole pair at the corner frequency of most filters causes problems with peaking at that frequency. This is in consistent with the dynamic range limitation, because
Figure 3.1: $20 \log_{10} \left| \frac{H(j\omega_2) - H(j\omega_1)}{j\omega_2 - j\omega_1} \right|$ versus input signal frequency $\omega_2$ and output noise frequency $\omega_1$ for a 5th order, 0.01 dB ripple, -49 dB stopband Elliptic and a 7th order 0.2 dB ripple Chebyshev (these have the same transition bandwidth). Sensitivity depends on the value along the diagonal line, and blocking performance depends on the value in the boxes. Isopleths (colours) are 2 dB wide.
\[ |H(j\omega_1) - H(j\omega_2)|/j(\omega_2 - \omega_1) \] is usually largest for \( \omega_1 \) and \( \omega_2 \) near the corner frequency - see the red regions in fig. 3.1. (Note that the change in transfer function phase, i.e. group delay, contributes as well as the change in transfer function magnitude.)

As filter designers know the significance of pole Q, it is reasonable to ask whether the dynamic range limitation makes any different predictions. Now some authors (Premoli 1973) (Kim 1992) have suggested replacing the high-Q pole pair at a filter’s corner frequency by multiple lower-Q poles to decrease the sensitivity. Considerations based on pole Q suggest this is worthwhile, but the dynamic range limitation rules out much improvement from such techniques if the transfer function remains similar. The idea does not seem to be used much in practice.

### 3.3 REALISATION

Having chosen the transfer function, we move on to the realisation.

#### 3.3.1 Unsuitability of the frequency-weighted input-normal realisation

Suppose we wish to use the frequency-weighted input-normal realisation, which was shown in the last chapter to theoretically give optimal dynamic range. Typical weights might be

\[
\Phi(j\omega)^2 = \begin{cases} 
1, & |\omega| < 2 \\
10, & |\omega| \geq 2
\end{cases}
\]

(3.1)

assuming blockers are 10 dB larger and start at \( \omega = 2\text{rad/s} \), and

\[
\Psi(j\omega)^2 = \begin{cases} 
1, & |\omega| \leq 1 \\
0.1, & 1 < |\omega| < 2 \\
1, & |\omega| \geq 2
\end{cases}
\]

(3.2)

assuming that the ADC sample rate is three times the filter bandwidth, so noise above \( 2\text{rad/s} \) folds into the passband etc. (In practice, \( \Psi(j\omega)^2 = 1 \) everywhere gives almost identical results. Increasing \( \Psi(j\omega)^2 \) further may also decrease stopband sensitivity.)
The realisation which results is

\[
A = \begin{bmatrix}
-0.0240 & 0.4060 & -0.0253 & -0.0257 & -0.0053 \\
-0.6525 & -0.1956 & 0.4146 & 0.0719 & 0.0229 \\
-0.4458 & -0.9923 & -0.3734 & -0.3273 & -0.0675 \\
0.6989 & 0.6483 & 1.1722 & -0.4386 & -0.3526 \\
-1.0441 & -1.1410 & -1.2403 & 1.7061 & -1.0441
\end{bmatrix}
\]

\[
B = \begin{bmatrix}
-0.4618 \\
-0.4001 \\
-0.3821 \\
0.3083 \\
-0.4705
\end{bmatrix}
\]

\[
C = \begin{bmatrix}
-0.1759 & 0.3960 & -0.3366 & -0.1244 & -0.0287
\end{bmatrix}
\]

\[
D = \begin{bmatrix}
0
\end{bmatrix}
\]

The problems with this realisation should be apparent. The dense tap matrices will require lots of resistors. Moreover, small taps such as \(a_{11}\) require large resistors (100k\(\Omega\)), which will be physically large and hence have high parasitic capacitance. (Sometimes they can be rounded to zero without changing the transfer function much though.) Large taps such as \(a_{54}\) will heavily load the relevant integrator, requiring the unit resistance to be increased to prevent current limiting. This then results in increased thermal noise. Nevertheless, these problems should not be considered insurmountable – the author used this approach once in a 30 MHz filter. It is also possible to halve the number of non-zero coefficients (Li, Anderson, Gewers & Perkins 1992) or partition the system into lower-order sub-filters. (Partitioning is sub-optimal though.)

### 3.3.2 Merit of lossless realisations

As the frequency-weighted input-normal realisation was not very suitable, a more conventional realisation was used, namely a simulation of a doubly-terminated LC ladder. The state-space equations describing an LC ladder are known to have good sensitivity properties (Fettweis 1986) (Orchard et al. 1985). (These papers give a useful introduction
and refer back to most of the original work which dates from the 1960’s.) The passband sensitivity is low by the Orchard / Fettweis argument that changing the L or C values cannot allow the gain of a doubly-terminated LC network to exceed 0 dB. Hence at frequencies where the gain is already close to 0 dB, the first-order sensitivity must be low to prevent the gain exceeding 0 dB with a small parameter variation. The stopband sensitivity of an LC ladder is good because the ladder structure does not require two nearly equal quantities to cancel to achieve the stopband rejection. The stopband attenuation is achieved progressively along the ladder.

We can quantitatively compare the frequency-weighted input-normal realisation and the LC ladder simulation using tr $W^\Psi$ with the diagonal elements $k_{ii}^\phi$ (the powers in the states for input spectrum $|\Phi(j\omega)|^2$) of $K^\phi$ scaled to unity. For the weights (3.1) and (3.2) and the Elliptic transfer function used, the result is tr $W^\Psi = 1.59$ for the frequency-weighted input-normal realisation and tr $W^\Psi = 1.02$ for the LC ladder. The LC ladder is actually better! This may appear to contradict the optimum nature of the frequency-weighted input-normal realisation. However, it actually results from using a slightly different norm to that in which the frequency-weighted input-normal realisation is optimal. The LC ladder spreads the signal energy across the states better than the frequency-weighted input-normal realisation. An “equal-wordlength optimal realisation” exists (Mullis & Roberts 1976) which is better again than the LC ladder, but it is not discussed here as it has little connection to the Hankel singular value theory, requires a dense tap matrix, and is only a little better.

In a lot of cases, an LC-ladder-based realisation achieves an output noise tr $W^\Psi$ close to $\frac{2}{3} tr K^\phi W^\Psi$. The author does not know of any theoretical justification for this. However, it is very useful to know that the dynamic range achieved by a practical realisation, namely the simulation of an LC ladder, usually correlates closely with the norm tr $K^\phi W^\Psi$ which can be calculated from the transfer function alone. This consideration was the basis of the author’s earlier paper (Harrison & Weste 2001).
3.3.3 Lossless Realisation

Section 3.2 showed that transfer functions with finite stopband zeroes (notably low-ripple Elliptic ones) were superior, particularly with respect to sensitivity at the passband edge. However, this complicates the state-space realisation. The fifth order LC ladder has five capacitors and two inductors, so a state consisting of all capacitor voltages and inductor currents is not minimal. This problem was solved by using the descriptor state-space equations (which have an extra matrix in front of $\frac{dx}{dt}$) in place of the state-space equations.

That is, the floating capacitors ($C_2$ and $C_4$) in the Elliptic LC ladder prototype were realised as floating capacitors in the opamp-RC realisation. The descriptor state-space equations are

$$\frac{dx}{dt} = Fx + Gu$$  \hspace{1cm} (3.3)

$$y = Hx.$$  \hspace{1cm} (3.4)

The state is taken as

$$x = \begin{bmatrix} v_{C_1} & i_{L_2} & v_{C_3} & i_{L_4} & v_{C_5} \end{bmatrix}^T,$$
from which is is straightforward to analyse the circuit and derive that

\[
C = \begin{bmatrix}
C_1 + C_2 & 0 & -C_2 & 0 & 0 \\
0 & L_2 & 0 & 0 & 0 \\
-C_2 & 0 & C_2 + C_3 + C_4 & 0 & -C_4 \\
0 & 0 & 0 & L_4 & 0 \\
0 & 0 & -C_4 & 0 & C_4 + C_5
\end{bmatrix}
\]

\[
F = \begin{bmatrix}
-\frac{1}{R_1} & -1 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 \\
0 & 1 & 0 & -1 & 0 \\
0 & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 1 & -\frac{1}{R_5}
\end{bmatrix}
\]

\[
G = \begin{bmatrix}
2 & 0 & 0 & 0 & 0
\end{bmatrix}^T
\]

\[
H = \begin{bmatrix}
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

A similar matrix-based approach for gm-C filters is in Greer, Henderson, Ping & Sewell (1991). As the matrix C is diagonally dominant, the dynamic range properties of the descriptor equations are similar to the true state space equations. True state-space equations can be obtained by algebraic manipulations, but the tap matrix then becomes dense. (This is how \(K_0 W^q\) was calculated in section 3.3.2.) The LC element values were taken from the tables in (Zverev 1967).

An important consideration is that \(F\) and \(H\) only have unit elements. (\(R_1\) and \(R_5\) are 1 \(\Omega\).) The coefficient of two in \(G\) is easily realised by two parallel elements. This means that the filter can be tuned by adjusting a unit element.

### 3.3.4 Filter taps - practical details

Each element \(C_{ij}\) of \(C\) corresponds to two capacitors, and each element of \(F, G\) and \(H\) corresponds to two switchable unit resistors - see fig. 3.3. (\(G\) uses two pairs of parallel unit resistors as the tap coefficient is two).

The digital process used does not have metal-insulator-metal capacitors, so routing
metal was used for capacitors. Metal layers 3, 4, 5 and 6 were used, giving a capacitance of $3 \times 40 \text{aF}/(\mu\text{m})^2 = 120 \text{aF}/(\mu\text{m})^2$. The bottom-plate capacitance is around $10 \text{aF}/(\mu\text{m})^2$. Capacitors were built from a unit cell of 15 fF. The diagonal elements $C_{ii}$ are 150 fF to 360 fF, and the floating capacitors $C_{ij}$ to realise the stopband zeroes are 15 fF and 45 fF. Some rounding of tap values is necessary to avoid too small a unit cell. The frequency response with rounded tap values was checked in Matlab. In general, lossless simulations can tolerate a surprising amount of rounding - a Matlab Monte Carlo simulation with 5% capacitance mismatch still had acceptable frequency response. This knowledge is helpful for layout where a decision often has to be made between low parasitics and optimum matching. Given the tolerance of mismatch, here the emphasis was on low parasitics so common-centroid geometries were not used.

The filter is tuned by switching binary-weighted resistors. The resistors are un-silicided polysilicon. The most significant bit (MSB) is a resistor with a single finger $R = 3.4 \text{k}\Omega$, and proceeding towards the least significant bit (LSB) the resistors are $R, 2R, 4R, \cdots$. A practical filter would have five or six bits of tuning. The parasitic capacitance of the LSB’s needs to be considered, but it is not prohibitive if the resistor width is kept below about 1 μm. In this process the polysilicon matching is fairly good (around 1% - μm), so mismatch is not a significant problem. If matching was worse, the MSB could be constructed from a series-parallel combination rather than a single finger. In the filter fabricated, only three bits of tuning were included due to pin count restrictions for probing.

There is some subtlety in implementing switched tuning. Matching favours one resistor finger or capacitor cell being an LSB rather than an MSB. With capacitors, this is practical. Tuning the capacitors also sometimes allows the common-mode level of the switches to be at ground, lowering the on-resistance. However, in this design the capacitors vary from 15 fF to 360 fF. The LSB’s of a 15 fF capacitor are too small to tune. (Behbahani, Tan, Karimi-Sanjaani, Roithmeier & Abidi (2000) avoid this problem by not tuning the small capacitors which realise the stopband zeroes, with the consequence that the stopband zeroes are at the wrong frequency at tuning extremes.) If resistors are tuned and it is desired that one finger is an LSB rather than an MSB, then the switched resistors need to be in series rather than parallel. However, switches in series need to be wider, and the parasitic
capacitance is unacceptable. Sometimes there is merit in tuning both capacitors and resistors, with the MSB’s tuned on one and the LSB’s tuned on the other (Behbahani et al. 2000).

3.3.5 Replica oscillator

To tune the filter to the correct cutoff frequency, the RC time constant must be measurable. This is commonly done using a replica oscillator which uses identical resistor and capacitor elements to the main filter (Nauta 1993) (Tsividis & Gopinathan 1994). Probe pad count and schedule restrictions prevented a replica oscillator being fabricated, but the circuit in fig. 3.5 performed well in simulation.

Transistors are used in place of diodes to control the oscillation amplitude because diodes would give too high an amplitude. The replica oscillator cutoff frequency correlated with the filter −3 dB frequency to 2% in Monte-Carlo simulations.
Unit resistor elements with switched tuning

Figure 3.4: Filter layout

Figure 3.5: Simulated replica oscillator
3.4 OPAMP DESIGN

3.4.1 Choice of opamps as the active element

Feedback circuits such as opamp stages have the advantage that the active element does not have to be as linear (before feedback). This should allow greater signal swing for a given distortion.

It is conventionally argued that the problem with opamps is that they do not have sufficient bandwidth (Toumazou, Lidgey & Haigh 1990, p. 347) (Nauta 1993, p. 7). More cautious authors (Tsividis & Gopinathan 1994, p. 201) admit that this argument may not be valid. The author does not know whether opamps are slower than transconductors, and whether voltage-mode circuits are slower than current-mode circuits. The argument that transconductors and current-mode circuits are faster because they avoid internal nodes with long time constants seems spurious. Presumably it stems from the method of open-circuit time constants (Gray & Searle 1969, p. 531). Unfortunately, for Miller-compensated and feedforward opamps, this method seems invalid. For a feedforward opamp, the feedforward path skips the long time constants. For a Miller-compensated opamp, the poles split and hence pole-by-pole calculations are inappropriate.

The argument in (Tsividis & Gopinathan 1994, p. 201) that opamp-RC filters are slow because CMOS opamps which drive resistors are slow is more plausible. However, the ultimate limitations on opamp gain-bandwidth may turn out to depend more on load capacitance rather than load resistance. The usable opamp bandwidth is some fraction of its unity-gain bandwidth. (By the Bode gain-phase relationships, the fraction depends on the phase lag.) At frequencies close to unity gain, the impedance of the capacitors in a filter is significantly lower than the impedance of the resistors. (For the filter described here, at the 3 GHz unity-gain frequency the integrator capacitance of 250 fF is \(- j 210 \Omega\). The resistive load on each opamp is around 1 k\(\Omega\).) It is therefore the load capacitance rather than resistance which limits the unity-gain frequency, and hence also the operating bandwidth.
3.4.2 Fundamental limitations on opamp bandwidth

What opamp bandwidth is possible? This problem can be split into two parts:

1. What opamp unity-gain frequency is possible?

2. At what fraction of unity-gain frequency can reasonable loop gain (30–40 dB) be obtained?

Unity-gain frequency

Using a simple small-signal model of a transistor which consists of only \( g_m, C_{gs} \) and \( C_{ds} \), the unity-gain frequency \( f_0 \) of cascaded identical common-source stages is

\[
f_0 = \frac{1}{2\pi} \frac{g_m}{C_{gs} + C_{ds}} = f_T \frac{C_{gs}}{C_{gs} + C_{ds}} \approx \frac{f_T}{2}.
\]

(Deep-submicron CMOS processes have a notably large \( C_{ds} \).) This is a fairly accurate upper limit on opamp unity-gain frequency. A fundamental limit can be calculated using the theory of Thornton (1957) and Glasser (1988). If perfect inductors and available, the maximum unity-power-gain frequency is \( f_{max} \). Without inductors, the techniques in (Glasser 1988) can be used to calculate that oscillation and hence amplification is only possible up to around \( f_T/2 \) for a typical deep-submicron CMOS transistor. Capacitance from pMOS current sources and routing reduce this a little further.

If the transistors could be biased for maximum \( f_T \) (30–40 GHz for a 0.18\( \mu \)m process), very impressive results would be obtained. In deep-submicron CMOS analog circuits, bias for maximum \( f_T \) is rarely possible, for two reasons:

- The \( V_{gs} \) (0.8 V) and \( V_{ds} \) (> 0.5 V) required compromises signal swing or tolerance of threshold, temperature and \( V_{DD} \) variations.

- The current consumed is excessive. For optimum bandwidth, transistor capacitance needs to dominate load capacitance. This requires wide transistors. A transistor with 100\( \mu \)m width operating at current density of 100\( \mu \)A/\( \mu \)m consumes 10 mA.
In this circuit, the transistor current density was limited to $10 \mu A/\mu m$. At this current density, the $f_T$ is around 10 GHz. Therefore cascaded identical common-source stages have a unity-gain frequency around 5 GHz. The opamp unity-gain frequency achieved was 3 GHz.

**Bandwidth with reasonable loop gain**

There is a bandwidth penalty in using feedback. With conventional one-pole gain rolloff ($-20 \text{dB/decade}$), the bandwidth penalty is obviously severe – a factor of 100 in bandwidth for 40dB loop gain. For switched-capacitor circuits, one dominant pole is necessary for fast settling. However, for continuous-time circuits it is usually bandwidth rather than settling time which is important. Acceptable loop gain can be achieved over a greater bandwidths by using a faster gain rolloff, such as around $-30 \text{dB/decade}$ (Bode 1945). Larger gain slope of course reduces the phase margin, but $45^\circ$ is usually quite sufficient. An approximation to this gain slope is two-pole compensation (Roberge 1975), where the gain falls off at a two-pole rate at low frequencies and a one-pole rate closer to the unity-gain frequency. (This is actually preferable to a constant $-30 \text{dB/decade}$ slope as the phase margin is improved.)

More feedback than this is possible, such as by using a conditionally stable amplifier. The results in section 2.10.1 are relevant here. The author guesses that 40dB of loop gain is possible at one fifth of the unity gain frequency. The results obtained from the procedure in section 2.10.1 depend heavily on the asymptotic rolloff of the gain and the tolerable peaking or out-of-band sensitivity. (For example, is an opamp which has a gain which falls to unity at 10 GHz but stays at 0.9 up to 1000 GHz possible?)

Fig. 3.6 gives a summary of the situation regarding unity-gain bandwidth and bandwidth for workable loop gain.

**3.4.3 Effect of finite opamp gain-bandwidth**

Given that opamp gain-bandwidth is limited, what opamp gain is actually required?

Consider a filter with transfer function $H(s)$ built from integrators with gain $k/s$. If integrator imperfections (notably finite opamp gain) result in an integrator transfer function $k/F(s)$, then the filter transfer function will be $H(F(s))$. This assumes that
Can trade-off DC gain
vs. bandwidth

Estimated ultimate limit
with practical sensitivity

Likely to be conditionally-stable

Figure 3.6: Achievable opamp gain vs. frequency.
imperfections in all integrators are the same. There will be some differences because
different integrators have different loads, but much of the variation will be shared.

There does not seem to be a good reference to this simple analysis in the literature. For
LC filters, the equivalent assumption, namely uniform element losses, is considered in
(Bode 1945) and (Blostein 1967). The author is presumably not the first person who has
had to re-invent this analysis, so he implores textbook writers to include useful
approximations like this rather than rigorous sensitivity analyses which offer less insight
and are superseded by Monte-Carlo simulations.

For an opamp-RC integrator, if the opamp has finite gain $A(s)$ but no other
imperfections, it can be easily calculated that

$$F(s) = s + \frac{1}{A(s)}(s + k).$$  \hspace{1cm} (3.5)

The opamp used here (and most others for that matter) has a phase shift $\angle A(s)$ of around
$-90^\circ$ near the filter cutoff. $(s + k)$ is around $+45^\circ$. Together, these result in the second
term in (3.5) having a phase shift around $+135^\circ$, moving $F(j\omega)$ to the left half plane. This
causes peaking as $F(j\omega)$ is closer to the filter poles. The peaking is most pronounced
around the cutoff frequency as the high-Q pole pair there is closest to the $j\omega$ axis.

The amount of peaking is straight-forward to calculate. A reasonable approximation to
the opamp gain $A(s)$ was chosen, and $A(j\omega)$, $F(j\omega)$ and $H(F(j\omega))$ were calculated using
Matlab. The result is that an opamp gain $|A(j\omega)|$ of 32 dB at the cutoff frequency results in
1 dB of peaking. A gain of 28 dB results in 2 dB of peaking. In practice, a few decibels more
gain is required because opamp input capacitance and routing parasitic capacitance on the
opamp inputs cut the effective loop gain. In this filter the reduction in loop gain is around
3 dB. Therefore 35 dB opamp gain is required for 1 dB of peaking, and 31 dB for 2 dB of
peaking. The actual opamp gain achieved was 35 dB at 200MHz and 30 dB at 350 MHz.

The tolerable deviations in the transfer function obviously depend on the application.
For a radio receiver, worst-case passband ripple of 2–3dB is often specified. To allow for
process variation, nominal peaking should be kept to about 1 dB, ie. 35 dB opamp gain.
The opamp used achieves this at 200 MHz. At 350 MHz cutoff frequency, $Q$-tuning would be
required. It is often said that 40 dB opamp gain is a good target, but with a well-rounded
transfer function like that used here slightly less suffices.

Another way of doing the same analysis is to simulate an integrator to get \( F(j\omega) \) (Nauta 1993). The integrator phase shift (fig. 3.7) is usually more significant than the integrator gain as gain errors just cause slight tuning errors and will be tuned out by the frequency tuning. (This may not be so true for disk-drive read channels where filter group delay variations are a problem.) Acceptable phase error is often around 1–2°. This has the advantage that opamp input capacitance is considered, although routing capacitance can still cause unpleasant surprises.

A useful interpretation of this is that an integrator with phase lag less than 90° is like a lossy inductor or capacitor in the LC ladder prototype. Dissipated energy results in passband droop near the corner frequency. There is a quantitative theory of these effects (Blostein 1967) (Kishi & Nakazawa 1963) (Kishi & Kida 1967) (Livsic 1973) (Harrison & Weste 2001) involving stored energy and group delay which makes similar predictions to those from the dynamic range limitation. (The dynamic range limitation is superior in that it applies to all transfer functions and better considers blocking dynamic range though.) For the converse situation of integrator phase lag greater than 90°, the storage elements generate rather than dissipate energy and the response peaks rather than droops. By contrast, integrator gain errors preserve the passivity of the filter and cause less amplitude deviation.

3.4.4 Gain stage principles

Close to optimal performance should be possible with simple common-source stages. There is then the issue of how to connect multiple stages together.

In the past, this issue was complicated by the fact that common-source stages have negative gain, and common-gate and source follower stages have positive gain. It was often necessary to choose common-gate or source-follower stages simply because a positive gain was required, not because there was any bandwidth or signal swing advantage. An example of this is in two-stage single-ended feedback amplifiers.
Figure 3.7: Integrator phase shift – Monte Carlo simulation. 20 runs at each of −25°C, 25°C and 125°C.
With the rise of fully-differential circuitry in the 1990’s, this problem is less important. The sense of the gain can be changed by swapping the positive and negative inputs (or outputs). The problem is not entirely eliminated, because occasionally common-mode latch-up or oscillation problems arise. (This is latch-up in the opamp not CMOS sense of the word.)

(The complexity of transistor-level analog circuit design seems to be largely a consequence of the unavailability of ideal transformers and ideal blocking capacitors. The theory of Thornton (Thornton et al. 1966) (Thornton 1957) on fundamental limitations on transistor circuits is very interesting in this context, and seems to be applicable to the current-mode versus voltage-mode debate. The present era seems to be the Baroque period of circuit design.)

Common-mode feedback

The difficulty with fully-differential circuits is that common-mode feedback is required (Duque-Carrillo 1993). Building a good common-mode feedback circuit is harder than might be expected. Switched-capacitor common-mode feedback is good, but it cannot be used in a continuous-time filter. The following issues are significant:

- When the opamp is operating in a circuit in which the input common-mode level is set by the output common-mode level, latch-up with both inputs and both outputs stuck at a supply rail can occur. Common-mode feedback circuits that might be expected to work often cannot recover from this condition. The circuit in (Banu, Khoury & Tsividis 1988) is one which does avoid this problem.

- The common-mode feedback loop may need compensation. This is particularly an issue in two-stage opamps. What bandwidth should the common-mode loop have?

- The gain of the common-mode feedback loop can be a fine trade-off between achieving sufficiently precise control of the common-mode level, and not degrading the differential linearity. This is particularly a problem if a transistor circuit with limited linearity is used to average the two outputs to derive the common-mode level.
The author has reached the conclusion that the best way to compensate a two-stage opamp is for each stage to have its own common-mode feedback. The virtues of local (single-stage) feedback are that it does not need compensation and is usually globally stable. The virtue of global (multiple-stage) feedback is that the loop gain is higher so the error is smaller. In common-mode feedback circuits, the common-mode level usually only needs to be set to within 100 mV or so. This does not require the precision of multiple-stage feedback, so the virtues of single-stage feedback are more important. Therefore two local common-mode feedback loops should be used instead of one feedback loop around two stages.

The nMOS-nMOS opamp

A two-stage opamp traditionally has a pMOS input stage and an nMOS output stage. The two stages need to use opposite types of transistors to achieve good input common-mode range and output swing (Cherry & Hooper 1968, p. 286). The pMOS / nMOS combination is preferred over nMOS / pMOS as bandwidth of the output stage is more critical, and pMOS transistors often have lower flicker noise.

However, as transistor threshold voltage $V_T$ approaches $V_{DD}/2$, an nMOS input stage and an nMOS output stage becomes more viable. The maximum signal swing of cascaded identical stages like this is $2V_T$ peak-to-peak per side. Obviously, with a 15 V or 5 V power supply this is sub-optimal. But with $V_{DD}=1.8$ V and $V_T=0.5$ V, the signal swing achieved is over half of rail-to-rail. For high speed circuits, this is important because of the higher $f_T$ of nMOS transistors. It also simplifies design somewhat, particularly when implementing feedforward.

3.4.5 Compensation

A two-stage opamp is usually used for opamp-RC filters, as single-stage opamps do not achieve sufficient gain with resistive loading. The usual method of compensating a two-stage opamp is Miller compensation. This is a good compensation technique – unity gain frequencies of a significant fraction of $f_T$ are possible. It also has the virtue that the
Miller capacitor provides additional feedback around the output stage. However, intuitively Miller compensation can be expected to be sub-optimal for two reasons. Firstly, a one-pole compensation with a gain slope of \(-20\,\text{dB/decade}\) is not optimal loop-shaping. The loop gain at moderate frequencies can be increased by a faster gain rolloff, say \(-30\,\text{dB/decade}\) (Bode 1945). This can be approximately achieved by using two-pole compensation (Roberge 1975). Secondly, the energy stored in the Miller capacitor might better be used by an active device to generate signal power. Opamps which use the load capacitance for compensation may have advantages.

The feedforward compensation used in the filter opamp does not have these drawbacks. The gain declines at a two-pole rate (from the main signal path) at low frequencies, and a one-pole rate (from the feed-forward path) at high frequencies. This is better shaping of the sensitivity versus frequency, and it avoids compensation capacitors. How this is achieved is easiest to see from the circuit.
3.4.6 Opamp circuit

As might be expected considering the discussion above, the opamp (fig. 3.8) is constructed from nMOS differential pairs with local common-mode feedback. The main low-frequency signal path is through the two differential pairs on the right. The differential pair on the left provides the high-frequency feedforward path which pulls the phase lag back to 90° at high frequencies to provide acceptable phase margin.

The stages use simple local common-mode feedback circuits composed of pMOS current source transistors with resistors or long transistors to derive the common-mode output. Long, narrow (0.9/18) transistors are used for the input stage. These provide a higher resistance and lower capacitance than resistors would. Resistors are used for the output stage though as they are more linear.

For simplicity, transistors in all three differential pairs run at the same current density, 10 μA/μm. The current density is limited by the need to provide sufficient $V_{ds}$ across the tail current sources at low temperature and slow (high $V_T$) process corner. For typical transistors, the signal common-mode level is around 1.1 V, and the sources of the differential pairs sit at 0.4 V. In future designs, a more sophisticated CMFB which sets the common-mode level at $V_{gs} + V_{sat}$ rather than $V_{DD} - V_{gs}$ could be used. This would reduce the minimum power supply voltage from $2V_{gs} + V_{sat}$ to $V_{gs} + 2V_{sat}$, allowing higher transistor current density for the same $V_{DD}$ and hence higher bandwidth.

3.5 MEASURED PERFORMANCE

3.5.1 Frequency response

The measured frequency response (fig. 3.10) has two main imperfections. One is the peaking at high cutoff frequencies due to finite opamp gain. This is in agreement with the theory in section 3.4.3. The other effect, which was not present in simulation, is some passband droop. There are a number of imperfections which could cause this, such as capacitor losses, routing resistance and lower than expected opamp DC gain. Routing
Figure 3.9: Opamp Bode plot. Load 300 fF in parallel with 1 kΩ

resistance seems the most likely. It was not possible to include parasitic resistance in extracted simulations, as it adds too many nodes to the circuit and the simulator crashes when its matrix exceeds 4 Gbyte in size. (This problem is largely caused by a limitation of the extraction program, Cadence Diva, which leads to a node being created at every via.) The routes between the opamps and capacitors have an estimated resistance of 80 Ω, which is sufficient to reduce the integrator phase lag by 2° and cause the observed passband droop.

Overall, the frequency response is sufficiently controlled for practical applications such as radio receivers for cutoff frequencies up to 200 MHz. A Monte-Carlo simulation at 180 MHz nominal cutoff frequency is show in fig. 3.11. At higher cutoff frequencies (350 MHz), Q-tuning would be required to control the peaking due to finite opamp gain.
Figure 3.10: Measured frequency response over the tuning range. Some of the stopbands are omitted for clarity. Tuning 001 (40 MHz cutoff) to 111 (350 MHz cutoff).
Figure 3.11: Monte-Carlo simulation of the extracted circuit over expected process and mismatch variations. 20 runs at each of −25°C, 25°C and 125°C. Tuning at nominal cutoff frequency of 180MHz.
Figure 3.12: Measured, simulated and prototype frequency responses (350 MHz responses are offset by +10 dB)
3.5.2 Dynamic range

The maximum signal swing is 0.50 Vp-p differential for ~40 dB total harmonic distortion (THD). This is with 150 MHz and 160 MHz input tones and 350 MHz cutoff frequency. The distortion does not vary much with input frequency, although it degrades a little if intermodulation products fall near the cutoff frequency, and improves a little at low frequencies. The worst distortion products are at 460 MHz and 470 MHz (around −45 dBc); the IM3’s at 140 MHz and 170 MHz are around −50 dBc. Measured and simulated distortion are similar. The table below summarises the performance.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>14 mA @ 1.8 V</td>
</tr>
<tr>
<td>Gain</td>
<td>0 dB</td>
</tr>
<tr>
<td>Tuning range (−3 dB cutoff)</td>
<td>40 MHz – 350 MHz</td>
</tr>
<tr>
<td>Inband IP3 (150,160 MHz)</td>
<td>−13 dBVRMS</td>
</tr>
<tr>
<td>Out-of-band IP3 (450,650 MHz)</td>
<td>−10 dBVRMS</td>
</tr>
<tr>
<td>Max. input for −40 dB THD</td>
<td>0.50 Vp-p</td>
</tr>
<tr>
<td>Average passband noise</td>
<td>24 nV/√Hz</td>
</tr>
<tr>
<td>Dynamic Range (−40 dB THD)</td>
<td>52 dB</td>
</tr>
<tr>
<td>Process</td>
<td>digital CMOS 0.18 µm</td>
</tr>
<tr>
<td>Area</td>
<td>0.5 mm²</td>
</tr>
</tbody>
</table>

A slightly higher signal swing could perhaps have been hoped for – as discussed previously the voltage swing is limited to around 2Vp peak-to-peak per side, ie. 2.0 Vp-p differential. This filter is limited more by output stage current than voltage swing; at 0.50 Vp-p the devices in the opamp output stages swing from 0.4 mA to 1.1 mA drain current. The opamps clip from lack of output stage current when the output is around 1.0 Vp-p. There is also some contribution to the distortion from the switched tuning transistors.

The average noise over the passband is 24 nV/√Hz. The impedance levels were kept as low as possible (a little too low in fact, because the opamps clip from lack of output stage
current) to minimise thermal noise. The resulting dynamic range is 52 dB (for −40 dB THD).

The table below compares these results to others published for filters with cutoff frequency > 100 MHz. “Swing” is differential signal swing for −40 dB THD. “Φ_error” is the maximum integrator phase error in Monte Carlo simulations, for a frequency of 200 MHz.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Topology</th>
<th>$V_{DD}$</th>
<th>Swing</th>
<th>Φ_error</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>This work</td>
<td>opamp-RC</td>
<td>1.8 V</td>
<td>0.50Vp-p</td>
<td>1.2°</td>
<td>24 nV/$\sqrt{Hz}$ @ 25 mW</td>
</tr>
<tr>
<td>(Pavan et al. 2000)</td>
<td>gm-C</td>
<td>3.3 V</td>
<td>0.38Vp-p</td>
<td>?</td>
<td>14 nV/$\sqrt{Hz}$ @ 70 mW</td>
</tr>
<tr>
<td>(Gopinathan et al. 1999)</td>
<td>opamp-gm-C</td>
<td>2.5 V</td>
<td>0.46Vp-p</td>
<td>0.7°</td>
<td>?</td>
</tr>
<tr>
<td>(Rao et al. 1999)</td>
<td>gm-C</td>
<td>3 V</td>
<td>0.20Vp-p</td>
<td>1.5°</td>
<td>?</td>
</tr>
</tbody>
</table>

It is difficult to make a fair comparison considering the different applications, but the signal swing and dynamic range of the opamp-RC approach appears superior for a given $V_{DD}$. The noise in (Pavan et al. 2000) is identical to the filter presented here after scaling for power consumption.

These results are all without Q-tuning. There are also circuits which use Q-tuning, such as (Deng & Lee 2001) and (Nauta 1993). A signal swing of 0.3 Vp-p at 1 V $V_{DD}$ is achieved in (Deng & Lee 2001) using a current-mode approach. However, the current amplifier in (Deng & Lee 2001) still requires $V_{DD} > V_{gs} + 2V_{sat}$. It is only by operating at lower current densities and using Q-tuning to correct for the finite amplifier bandwidth (only 1.2 GHz) that 1 V operation is possible. (There also seems to be no mention of threshold variations.) Simple differential pairs with switched-resistor source degeneration (to control $g_m$) may also be attractive if Q-tuning can be used – an impressive dynamic range (albeit at lower frequencies and higher $V_{DD}$) is achieved in (Behbahani et al. 2000). Nevertheless, opamp-based high-feedback circuits seem the best possibility for achieving close to rail-to-rail signal swings with low distortion.

### 3.6 ADDITIONAL SIMULATIONS

Some additional simulations were performed later to see what bandwidth and dynamic range could be achieved by the opamp-RC topology. The filter capacitors were kept
The number on each stage (right) is the number of differential pairs in parallel.

unchanged, and the unit resistors decreased to allow tuning up to 1 GHz. More opamp power supply current was required to drive the lower impedances. Satisfactory opamp gain at 1 GHz was achieved with a factor of six increase in power supply current – this is not too bad considering that the impedances are one third of what they were. A remarkable opamp unity-gain frequency, 10 GHz, was achieved (fig. 3.14).

The opamp (fig. 3.13) has three stages with two feedforward paths. This increases the gain at low frequencies, with little expense in unity-gain bandwidth. The phase margin (fig. 3.14) is reduced to 20-30°. This may seem marginal, but the feedback path has around 20° lead, which increases the loop phase margin to 40°. The opamp phase margin also does not vary much under process variations – it is closely controlled by the relative strengths of the feedforward paths. Inadequate phase margin manifests itself as stopband peaking around 5-10 GHz; fig. 3.15 shows this is acceptable.

It is interesting to note that even if the phase lag exceeded 180° at lower frequencies, this opamp architecture will not necessarily be conditionally stable. This is because the main path will tend to clip from insufficient current before the feedforward path does. The reduced gain from the main signal path and unchanged gain from the feedforward path will reduce the phase lag. This offers interesting possibilities for opamps using very fast gain
Figure 3.14: Three-stage opamp Bode plot. Monte-Carlo simulation over expected process and mismatch variations. 20 runs at each of $-25^\circ\text{C}$, $25^\circ\text{C}$ and $125^\circ\text{C}$. Load 400 fF in parallel with 400 $\Omega$. 
roll-off to achieve high gain (> 60 dB) at gigahertz bandwidths.

The 10μA current source in each differential pair’s common-mode feedback (fig. 3.13, left) increases the common-mode output level by 0.2 V, to around 1.0 V. This increases the output swing. A fixed current was used here, but a feedback loop or replica structure could be used if a controlled output common-mode level was needed. A more sophisticated approach could also optimise the signal swing over $V_{DD}$ and process variations.

A Monte-Carlo simulation of filter frequency response when tuned to 800 MHz is shown in fig. 3.15. This is about the practical limit to the achievable bandwidth for acceptable frequency response without Q-tuning. The bandwidth achieved would actually be significantly higher (around 1.5 GHz) if not for the parasitic capacitances of the tuning switches. The tuning range was reduced to the minimum necessary to reliably cover process variations, so as to minimise the effect of switch parasitics. Zero-threshold transistors would be very helpful here as acceptable on-resistance could be achieved with narrower transistors. Switch width is a trade-off between capacitance and distortion from voltage across the switch.

The simulated performance is as follows:

<table>
<thead>
<tr>
<th>Specification</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>76 mA @ 1.8 V</td>
</tr>
<tr>
<td>Gain</td>
<td>0 dB</td>
</tr>
<tr>
<td>Tuning range (~3 dB cutoff)</td>
<td>500 MHz – 1 GHz</td>
</tr>
<tr>
<td>Inband IP3 (400,500MHz)</td>
<td>+8 dBV$_{RMS}$</td>
</tr>
<tr>
<td>Out-of-band IP3 (1.4,2.4 GHz)</td>
<td>+11 dBV$_{RMS}$</td>
</tr>
<tr>
<td>Max. input for ~40 dB THD</td>
<td>1.6 Vp-p</td>
</tr>
<tr>
<td>Average passband noise</td>
<td>18 nV$/\sqrt{Hz}$</td>
</tr>
<tr>
<td>Dynamic Range (~40 dB THD)</td>
<td>60 dB</td>
</tr>
<tr>
<td>Process</td>
<td>digital CMOS 0.18μm</td>
</tr>
</tbody>
</table>

A 10 dB higher signal swing and 20 dB higher IP3 than the fabricated 350 MHz filter has been achieved. This is a consequence of higher opamp gain and more careful choice of
Figure 3.15: Monte-Carlo simulation over expected process and mismatch variations. 20 runs at each of -25°C, 25°C and 125°C. Tuning at nominal cutoff frequency of 800 MHz.

opamp output stage current. The filter maintains low distortion (<-40dB THD) until it clips.

Hopefully this filter will be fabricated in the near future. Layout is likely to require more care. The integrator capacitors will probably have to be placed next to their respective opamps to prevent oscillation at 5-10 GHz. Routing capacitance will need to be minimised, and routing resistance extracted or carefully estimated. Nevertheless, the operation of the existing filter at 350 MHz suggests that the 800 MHz filter is practical. The parasitic capacitances assumed for the schematic have been found to be conservative for past designs.

3.7 CONCLUSIONS

Opamp-RC filters can equal or better the performance of gm-C filters at the high frequencies previously though to favour transconductor-based approaches. It seems almost certain that they will be competitive at lower frequencies too. Signal swings of a large
fraction of rail-to-rail are possible by using the feedback of opamps to minimise distortion, and switched polysilicon resistors to overcome the linearity limitations of transistor-based resistors or transconductors.

Opamp-RC filters can operate at frequencies up to around \( f_T/50 \). Two-pole feedforward compensation is superior to Miller compensation for wide bandwidths. It can be implemented using a simple and robust approach in which each differential pair has its own common-mode feedback. This approach has a certain logic to it in that the problems of stabilising common-mode feedback loops around multiple stages are avoided.

The dynamic range limitation presented in the previous chapter can aid the choice of filter transfer function. The insights are similar to those from pole Q considerations, but are more rigorous and are applicable to consideration of dynamic range as well as sensitivity. Plotting the kernel \( |[H(j\omega_1) - H(j\omega_2)]/j(\omega_2 - \omega_1)| \) of the Hankel operator is shown to be particularly useful for comparing transfer functions \( H(s) \).

Despite its provably good dynamic range properties, the frequency-weighted input-normal realisation is not well-suited to practical filter design as it has a dense tap matrix and wide spread of tap coefficients. State-space simulations of lossless systems such as LC ladders seem better in practice and even in theory.
4 ANALYTIC LIMITATIONS ON SIGMA-DELTA MODULATOR PERFORMANCE

4.1 INTRODUCTION

We now turn our attention from active filters to sigma-delta analog to digital converters. A model using an additive noise proved useful in chapter 2 for analysing dynamic range, and again it does here. In this case the additive noise is quantisation error, and the additive noise analysis is the well-known noise transfer function analysis. Like filters, sigma-delta modulators are interesting to study because they have a non-trivial relationship between overall performance (say, signal to noise ratio) and the performance of the sub-blocks (notably quantiser resolution).

In chapter 2 much of the attention was on deriving fundamental limitations on the achievable performance, so as to guide design effort. This will again be the case – the Bode integral results in this chapter are in some sense analogous to the dynamic range limitation. Sections 4.2 and 4.3 introduce sigma-delta modulators and the noise transfer function (NTF) analysis respectively. Section 4.4 derives fundamental limitations on the noise transfer function. There are two main results – one which applies to all sigma-delta modulators, and another which considers excess delay in the modulator feedback loop. The first has been discovered by other authors but we elaborate on it here; the second appears to be new. Section 4.5 considers links with other theory, including the work in chapter 2. Section 4.6 shows how these fundamental limitations can guide design.

4.2 THE SIGMA-DELTA MODULATOR

Sigma-delta modulators “work” by using a feedback loop to achieve resolution higher than that of their internal quantiser, at the expense of less signal bandwidth than the Nyquist rate. A popular structure is that of 1A in fig. 4.1, namely a loop filter $L(z)$ and a quantiser
in the forward path of a feedback loop. The feedback corrects for quantiser error in the same way that a feedback amplifier corrects for output stage distortion. This requires $|L(z)|$ to be large in-band in the same way that an opamp needs high open-loop gain at frequencies of interest.

In a sigma-delta ADC the loop filter is an analog filter, often switched-capacitor and hence basically discrete-time. However, different structures are possible – for instance, the sigma-delta ADC described in chapter 5 has the loop filter in the feedback path and does not use a conventional analog loop filter. The quantiser block actually consists of a flash ADC driving a DAC (digital to analog converter). However, these can usually be lumped together for theoretical analysis. The modulators in fig. 4.1 are single-loop modulators, but cascades of feedback loops (multiple-stage analog signal shaping or MASH modulators) are also commercially important. Multiple-loop modulators could be analysed using vector signals and a matrix loop filter, but that is not done here.

The term *sigma-delta* modulator is used here. The “sigma” and “delta” refer to
summing / integrating and subtracting / differentiating blocks in early modulators.

“Modulator” seems to date back to when the raw sigma-delta ADC output was considered for transmission along telephone lines and the like. It is used to describe both sigma-delta ADC’s and DAC’s. (Today’s sigma-delta DAC’s are not really sigma-delta de-modulators.) Others use “delta-sigma” in place of “sigma-delta” as it better describes the signal flow in early modulators. The author considers both terms largely meaningless with today’s loop filters, and wants to encourage a way of thinking in which the loop filter is a generic transfer function rather than an integrator or integrators. “Sigma-delta modulator” seems preferable to “oversampling ADC” as it excludes oversampling without feedback.

Despite sigma-delta modulators having a variety of different structures (see (Tewksbury & Hallock 1978) for some of the diversity), a consistent analysis using the so-called noise transfer function can be used. This can be done by adapting the noise transfer function analysis to the structure used, or by a loop transformation to a standard structure such as that of 1A in fig. 4.1. The loop transformation approach is mostly used here so that the fundamental limitations only need to be derived once. There is nothing unusual in such loop transformations; for example control systems engineers transform practically every problem they encounter into a standard feedback loop form.

4.2.1 Trade-off between resolution and oversampling ratio

As mentioned, sigma-delta ADC’s trade off bandwidth for resolution. The nature of this trade-off has different explanations (Candy & Temes 1992) (Norsworthy, Schreier & Temes 1997). Many authors consider it a reshaping of the quantisation noise spectrum away from frequencies of interest. In the terms of the filter analysis in chapter 2, we might say that we want a low noise gain from quantiser to output in the same way that we want a low noise gain $G(s)$ from state to output in a filter. There is more interpretation in section 4.4.2.

The oversampling ratio is defined as the Nyquist bandwidth (half the sample rate) divided by the useful signal bandwidth. It is obviously important for design to know what oversampling ratio (OSR) is required for what resolution. (“Resolution”, “signal to noise ratio (SNR)” and “dynamic range” here have similar meanings.)
The conventional approach to determining the oversampling ratio required for a given resolution is to try a range of loop filters $L(z)$. The sensible place to start is with first and second order loop filters, and then increase the loop filter order. For low-pass modulators with single-bit quantisers and discrete-time loop filters this research program seems reasonably complete. Good loop filters are known for high-order single-loop and low-order multiple loop (MASH) modulators (Schreier 1993) (Norsworthy et al. 1997). The results are usually presented as graphs of signal to (quantisation) noise ratio (SNR) versus OSR and loop filter order. Bandpass modulators, multi-bit quantisers and continuous-time loop filters are the subject of ongoing research in this context.

4.2.2 A subjective comment regarding use of “fundamental limitations”

There is no doubting the practical merit of the SNR versus OSR curves. Where they seem somewhat unsatisfactory is that they offer little insight into why a certain OSR is needed for a certain SNR. It seems reasonable to try to improve a circuit by understanding what is fundamentally limiting its performance, and what form the mathematical fundamental limitations take. For instance, the work in the previous chapter stemmed from knowing that the mathematical limitations on feedback amplifiers allowed far higher loop gains at wide bandwidths than were currently being achieved. The author would not have attempted to use opamps at such high frequencies if he did not know this. The derivations of such limitations are also useful to understand, as they often give useful design hints as to what factors lead to sub-optimal performance.

This approach of considering fundamental limitations seems to appeals strongly to some designers and design schools and not at all to others. It seems to be popular with Philips (Annema 1999) (Tellegen 1952) (Hurkx 1997) and particularly MIT researchers (Thornton et al. 1966) (Glasser 1988) (Mason 1954) and former students (Lee 1998). This is understandable, for two reasons. Firstly, the limitations usually make assumptions of limited validity. In this chapter the main assumption is that a linearised analysis (noise transfer functions and the like) is close enough, and that stability can be reasonably predicted from out-of-band noise transfer function. This largely ignores a lot of good work on nonlinear stability, such as (Baird & Fiez 1994), (Schreier, Goodson & Zhang 1997) and
(Farrell & Feely 1998).

Secondly, the form of the limitations is often unfamiliar. In this chapter they are integrals over frequency. These obviously appeal more to people who analyse feedback amplifiers using Bode plots and loop-shaping rather than root loci, poles and zeroes and polynomials. The author has a preference for techniques based on frequency response or impulse response rather than on realisations with a particular order and structure of tap coefficients. Mathematics went through a similar phase; the impression is that Newton thought of a function and its polynomial or power series as sort of the same thing (Stillwell 1989, p. 101) (Bourbaki 1994, p. 205). However, a more general concept of a function makes it easier to consider brick-wall ideal responses. These tend to be conceptually useful. It also avoids not finding the optimum because the wrong parameterisation has been chosen. Of course, to build a sigma-delta modulator you eventually have to choose a structure; the point is that there are advantages to doing this later rather than sooner. The practice of filter designers is better – the approximation (order, Butterworth vs. Chebyshev etc.) is separate from the realisation (integrators, inductors and capacitors). Sigma-delta modulator loop filters are filters (albeit with unusual sensitivity requirements), not some separate entity possessing a “life force” all their own.

4.3 NOISE TRANSFER FUNCTION ANALYSIS

In section 2.2 we calculated the dynamic range of state-space systems using an an additive noise model. We do a similar thing for sigma-delta modulators in this section. The sigma-delta modulator in fig. 4.1 1A is modelled as 1B. This structure plays a similar role to the one the state-space equations did in chapter 2. Quantisation noise is represented as an additive noise input $E(z)$.

For the modulator structure 1B in fig. 4.1, it is readily derived that

$$Y(z) = \frac{L(z)}{1 + L(z)} R(z) + \frac{1}{1 + L(z)} E(z)$$  \hspace{1cm} (4.1)

$$= G(z) R(z) + H(z) E(z)$$  \hspace{1cm} (4.2)

where $G(z)$ is called the signal transfer function, and $H(z)$ is called the noise transfer
function (NTF). The NTF is more important for deriving fundamental limitations, as the signal transfer function can be changed by an external pre-filter.

(Again we use the convention that upper-case quantities such as those in (4.2) are the Laplace transforms of the lower-case time-domain quantities of the same letter, here as labelled in fig. 4.1 1B. The notation in this chapter is similar to that in (Norsworthy et al. 1997), but with some control systems symbols. $H(z)$ is used as a generic transfer function in other chapters; in this chapter it will be used for NTF, although the limitations derived will sometimes be applicable to other quantities as well. $L(z)$ is the loop filter, but it includes any gain or delay in the quantiser or elsewhere in the loop – this is simpler for the long derivations which follow.)

At this point, no approximations have been made – the NTF and additive quantisation noise analysis is perfectly general and accurate. However, we know neither the quantisation noise at the output, the maximum signal input, nor whether the modulator will be stable. The output quantisation noise is the most straightforward. It is $H(z)E(z)$, so we need the noise added at the quantiser $E(z)$. Provided that the quantiser does not overload, the noise power can be calculated fairly accurately, as the quantisation noise has a peak amplitude of one half of an LSB (least significant bit of the quantiser). More problematic is what spectrum the quantisation noise $E(z)$ has.

4.3.1 Quantisation noise spectrum

The spectrum of quantisation noise is extensively discussed in (Gray 1990). It is well known that $E(z)$ is distinctly non-white in many practical modulators (Norsworthy et al. 1997). The worst problem is usually that $E(z)$ contains tones. This reduced spurii-free dynamic range (SFDR) is particularly a problem in audio applications, and radio systems susceptible to narrowband interference. (The effect of tones on radios varies a lot – a software radio or mobile telephone base-station will obviously have a problem if a tone knocks out a channel; on the other hand OFDM and CDMA systems depend more on high SNR than high SFDR.) Tones are usually worst in modulators with single-bit quantisers, low loop-filter order, no dither and little thermal noise, and for DC or other highly deterministic inputs.
Despite this, the linear NTF analysis is still useful even when tones are present. As the noise output is \( H(z)E(z) \) and the power of \( E(z) \) is limited (provided that the modulator does not overload), the amplitude of a large tone is still limited by the NTF \( H(z) \) at the tone frequency.

Fortunately, for the sigma-delta modulator presented in the next chapter the quantiser noise spectrum appears close to white in-band. This is presumably because it has a multi-bit quantiser and significant thermal noise.

### 4.3.2 Stability and maximum input signal

As with the state-space filter theory, maximum input signal must be considered as well as noise. Sigma-delta modulators have the rather peculiar feature that they can overload from their own quantisation noise. Moreover, they are often conditionally stable. Therefore, the maximum input signal problem cannot be decoupled from the noise problem like it can be for filters. A number of theories of sigma-delta modulator stability have been developed (Baird & Fiez 1994) (Hein & Zakhor 1993) (Schreier et al. 1997). Two are introduced here. The first, from (Schreier & Yang 1992) and (Kenney & Carley 1993), uses a 1-norm in the time domain and guarantees stability. The second uses an \( \infty \)-norm in the frequency domain. It does not guarantee stability, but gives a reliable and less conservative indication of it. It also fits very well with the integral limitations derived later in the chapter. The best expositions of it are (Norsworthy et al. 1997) and (Schreier 1993), but it was widely used before these.

\[ ||\hat{h}||_1 \] stability theory

The quantiser input is \((R(z) - E(z))G(z) = (R(z) - E(z))(1 - H(z))\). If we can guarantee that this does not overload the quantiser, the modulator will be stable. For a quantiser with \( n \) reconstruction levels with the usual equally-spaced decision and reconstruction levels, this can be shown (Kenney & Carley 1993) (Schreier & Yang 1992) to occur if

\[ ||\hat{h}||_1 < n + 1 - (n - 1)||r||_\infty \]

(4.3)
where \( \| r \|_\infty \) is the maximum input and

\[
\| h \|_1 = 1 + |h_1| + |h_2| + |h_3| + \cdots
\]
given the NTF impulse response

\[
H(z) = 1 + z^{-1}h_1 + z^{-2}h_2 + z^{-3}h_3 + \cdots.
\] (4.4)

Unfortunately, the sufficient condition for stability (4.3) is rarely satisfied by practical modulators with one-bit \((n = 2)\) quantisers. A maximum input of \( \| r \|_\infty = 0.5 \) is typical, requiring \( \| h \|_1 < 2.5 \). However, modulators with \( \| h \|_1 = 4.5 \) are usually stable (Schreier & Yang 1992), and offer much wider bandwidth. Interestingly, quite competitive bandwidth can be achieved with guaranteed stability - fig. 4.2 shows an NTF with \( \| h \|_1 = 2.5 \), but optimisation results in a strange FIR NTF with most of the taps equal to zero which would be near-impossible to realise in practice (except perhaps for sigma-delta DAC’s where it could be realised digitally). Using a more conventional NTF with \( \| h \|_1 = 2.5 \) results in a bandwidth of only one-tenth of that achieved with \( \| h \|_1 = 4.5 \).

\( \| H \|_\infty \) stability theory

When making practical stability predictions, another norm of the NTF is often more convenient and accurate than the 1-norm. This is the peak NTF in the frequency domain, \( \| H(e^{j\theta}) \|_\infty \). For high-order single-loop modulators with one-bit quantisers, it is empirically known that an \( \| H(e^{j\theta}) \|_\infty \) of 1.5 or 1.6 gives a good practical modulator (Schreier 1993) (Norsworthy et al. 1997) with a maximum input of around one half of the quantiser range. This frequency-domain condition will turn out to be very convenient when we consider integral limitations later in this chapter. Of course, as this design guideline is empirical the actual modulator being considered should be simulated to check the stability.

For multi-bit quantisers, the 1-norm approach guaranteeing stability is practical (Kenney & Carley 1993). Indeed, it is used in the following chapter. However, it is still conservative. Multi-bit quantisers are stable with NTF’s which have much higher \( \| H(e^{j\theta}) \|_\infty \) than those for one-bit quantisers. Many researchers in the field appear to be unaware of this fact or its importance. The literature rarely mentions it – (Ju & Vallancourt 1992) is the only study.
Figure 4.2: An NTF optimised with guaranteed stability, $H(z) = 1 - 1.1972z^{-1} + 0.2321z^{-8} - 0.0523z^{-26} + 0.0189z^{-37}$, and a more conventional good 5th order NTF with passband zeroes spread and an out-of-band gain of 1.6.
the author knows. A modulator with a 16-level (four-bit) quantiser and an out-of-band gain of \(|H(e^{j\theta})|_\infty = 5\) is stable with an input of 0.8 of the quantiser range (Ju & Vallancourt 1992), whereas (4.3) only guarantees its stability with inputs below 0.4 of the quantiser range. A modulator with \(|H(e^{j\theta})|_\infty = 5\) achieves four times the signal bandwidth of one with \(|H(e^{j\theta})|_\infty = 1.5\). The integral limitations derived shortly will show why.

Despite the fact that the \(|H(e^{j\theta})|_\infty\) criterion is fairly reliable and appears close to optimum for many practical cases, finding a better nonlinear stability criterion appears to be the “holy grail” of many researchers. This work is important, but more because it is likely to advance the art of nonlinear stability theory than because it will significantly improve practical sigma-delta modulators. The author’s opinion is that the existing state of the art is fairly close to fundamental limits. This will be discussed further later in the chapter.

Future developments are also likely to make a rigorous nonlinear stability analysis including both quantisation noise and maximum input less practical. Multi-bit modulators seem to be gaining in popularity for both ADC’s and DAC’s (with dynamic element matching). The linearised model works better for multi-bit modulators as the quantisation noise is whiter. (Checking for conditional stability with a Liapunov model is likely to remain important – the point is that the noise and stability questions become separate for many-level quantisers.) As speed and power consumption continue to be optimised, quantisers and integrators will become less ideal and hence require a more complicated mathematical model.

4.3.3 Applicability of the NTF analysis to the modulator investigated

The following chapter describes a sigma-delta ADC with a four-bit \((n = 16)\) quantiser. The linear analysis is particularly suitable for such a modulator. Indeed, even if a reliable and tight nonlinear stability test was found, it would probably be of little benefit. This is for the following reasons:

- The four-bit quantiser is likely to result in a quantisation noise spectrum which is closer to white than that of a one-bit quantiser.
• The modulator has such a wide bandwidth that thermal noise is probably large enough to have a dithering effect.

• The comparators in the quantiser have significant offset. An ideal quantiser cannot be assumed. Not much is lost by assuming that the additive quantisation noise is random.

• The modulator is absolutely stable and will recover from overloads which may result from an occasional bad combination of quantisation errors over successive cycles.

• The modulator is largely intended for radio applications such as wireless LAN’s. Such an input signal is unlikely to have a tight $\|r\|_\infty$. Overloads will occur from time to time. They are not catastrophic to the system.

All of these considerations favour a simple analysis based on average signal and quantisation noise powers. The NTF and signal transfer function (which is roughly 0 dB in-band) can be used to calculate an average power at the quantiser. Assuming a normally-distributed signal, an overload probability can be calculated. The NTF and input signal level can be chosen to make this acceptable. In the following chapter a 1-norm stability criterion was used, but a power (2-norm) criterion should work as well or better.

4.4 INTEGRAL LIMITATIONS

4.4.1 Links with Bode’s sensitivity integral

The NTF has the same expression $1/(1 + L(z))$ as the quantity called sensitivity in control theory. This is useful, because Bode (1945) derived a limitation on the integral of the log magnitude of sensitivity over frequency. This is the way the author originally derived the limitations which form the bulk of this chapter. Others have had the same idea independently (Nawrocki et al. 1999) and earlier (Gerzon & Craven 1989), although they seem not to have developed the links with linear prediction and the limitations regarding excess loop delay which are presented later in this chapter. $H(z)$ denotes both NTF and sensitivity in what follows. The equivalent continuous-time quantity will be $H(s) = 1/(1 + L(s))$. 

128
Bode’s integral is
\[ \int_{-\infty}^{\infty} \log |H(j\omega)| d\omega \geq 0. \quad (4.5) \]

That is, the sensitivity of a control system or feedback amplifier cannot be small at all frequencies. Distortion reduction can only be achieved at the expense of distortion increase at other frequencies. For a feedback amplifier, this practically manifests itself as distortion increase around the loop unity-gain frequency.

Bode’s result is for continuous time. The discrete-time equivalent is in (Sung & Hara 1988) and (Mohtadi 1990). It is
\[ \int_{-\pi}^{\pi} \log |H(e^{j\theta})| d\theta \geq 0. \quad (4.6) \]

The interpretation is the same – the sensitivity or NTF cannot be small at all frequencies. A small in-band NTF (eg. -100 dB in a 16 bit sigma-delta ADC) can only be offset by an NTF above 0dB out of band (eg. |H(e^{j\theta})| = 1.6, over a much wider bandwidth).

The situation where (4.6) is equal to zero is desirable, so it is worth explaining when this condition is met. It is met when the sensitivity is minimum-phase, ie. has no zeroes outside the unit circle. For (4.5), the integral is satisfied with equality when the sensitivity has no zeroes in the right-half plane; again this means that it is minimum-phase. (The integrals could be negative if the sensitivity had poles outside the unit circle or in the right half-plane, but this would mean that the system was closed-loop unstable and hence useless.) NTF zeroes outside the unit circle correspond to loop filter poles outside the unit circle, as \( H(z) = 1/(1 + L(z)) \). That is, the loop filter is open-loop unstable.

Open-loop-unstable loop filters have been proposed as a way of making a modulator chaotic so that it has less tonality (Schreier 1994). (This technique is viable despite a necessary degradation in NTF at some frequency, as (4.6) is only a little greater than zero if the loop filter poles are just outside the unit circle.)
Figure 4.3: The Bode integral states that the two shaded areas on this plot are equal. (The NTF shown is a 5th order inverse-Chebyshev with out-of-band gain around 1.5 (+3.5 dB). A fairly moderate in-band NTF (−40 dB) is chosen to make the diagram clearer.)

4.4.2 Interpretation

Eq. (4.6) has a good graphical description, namely that the areas above and below the x-axis on a plot of NTF in decibels versus frequency on a linear scale are the same. Fig. 4.3 provides a good example. (A chaotic modulator would have slightly larger area above the x-axis.)

This result is central to the author’s understanding of noise shaping. It is important to explain clearly why it is significant:

- It is possible to realise any NTF magnitude response which obeys the integral. The only limitation on an NTF is that it must be causal and have first sample of one (4.4). We shall see shortly why the first sample is always one and why it implies (4.6). It is a basic fact of filter theory that any magnitude response $|H(e^{j\omega})|$ is realisable by a causal, stable transfer function. The Bode gain-phase relationships (Bode 1945) imply that there is a certain associated minimum phase shift, but the phase of an NTF is
basically irrelevant. What matters is that any NTF magnitude response satisfying (4.6) can be realised. (This is meant with an engineering level of rigour; brick-wall filters can only be approximated arbitrarily closely, and presumably mathematicians have more pathological counter-examples.)

- The integral makes it clear that what OSR you need for a given in-band NTF comes down to two factors: how high the out-of-band NTF can be (whilst keeping acceptable stability), and how much bandwidth is “wasted” in the transition between the desired band (where the NTF is low) and the out-of-band region where the NTF is high ($|H(e^{j\omega})| = 1.6$ or whatever). Fig. 4.3 should make this clearer. A sigma-delta loop filter with high order and spread passband zeroes results in a modulator with less OSR for a given resolution solely because it wastes less transition bandwidth.

- The conventional explanation (Candy & Temes 1992, p. 2–3) of noise-shaping in terms of quantisation noise power on a linear scale is reasonable and correct, but gives no indication that it is impossible to have, say, no quantisation noise at all over half of the bandwidth. This is the strength of the Bode integral; the y-axis is a logarithmic scale on which a $-100$dB NTF is a lot different from a $-50$dB NTF, and an NTF is realisable if it satisfies the integral.

Overall, the point is that the result is useful for design, not just of theoretical interest.

Provisos and possible objections

Of course, the limitations of the linear NTF analysis should be kept in mind. For instance, whilst spreading the NTF zeroes should increase the useful bandwidth, sometimes it results in tones at frequencies between the zeroes and there is little advantage (Thoss, Li & Tenhunen 2000) (Vleugels 2000). Also, with a one-bit quantiser the loop filters $cL(z)$ (for a positive constant $c$) and $L(z)$ result in identical modulators. The single comparator in the quantiser does not measure magnitude. However, the NTF’s $1/(1 + cL(z))$ and $1/(1 + L(z))$ are different. The author’s opinion with respect to these and similar limitations is as follows:

- Design improvement often proceeds along the lines of Popper’s model of science (Popper 1959): come up with a hypothesis regarding design any way you can, and
then test the predictions it makes. The NTF analysis is good for making predictions. For example, the next chapter is based on the predictions that OSR can be decreased by increasing the out-of-band gain and using a multiple-bit quantiser to provide stability, and the prediction that an FIR loop filter will work as well as a conventional loop filter provided that it gives an appropriate NTF. Testing predictions about sigma-delta modulator performance is usually the easy part – just run a behavioral simulation with realistic non-idealities for 100 Monte-Carlo runs of 10000 cycles. It is not so important that the NTF analysis is sometimes of limited accuracy; what matters is that it points the designer in the correct direction about 80% of the time.

- The linearised theory is at least applicable to multiple-bit quantisers, which have a well-defined quantiser gain and less problems with tones.

There is another legitimate objection, with respect to treating loop filters $L(z)$ as just a filter transfer function to be approximated and realised. This is that the sensitivity and dynamic range requirements are quite different to those for a conventional filter. A filter like the anti-alias filter in the previous chapter might typically have 1–2dB tolerable passband error. Stopband errors of 10–20dB are quite acceptable provided that the stopband stays below the minimum stopband attenuation. Stopband phase shift is irrelevant. By contrast, a sigma-delta modulator loop filter can have large passband gain error provided that the gain stays sufficiently high. However, in the “stopband” of $L(z)$ where the NTF is high, 0.5dB gain errors or 20° phase errors are quite significant. These different sensitivity requirements are presumably why cascades of integrators are used. As an aside, the author suspects that ADC’s with continuous-time loop filters often achieve considerably less bandwidth than is possible because $L(z)$ is not sufficiently well controlled in the “stopband”. Dynamic range requirements are also peculiar – in-band distortion is reduced by feedback, and out-of-band distortion is mostly removed by the digital post-filter.

The main counter to this objection is that it still seems excessive to throw away ninety years of filter realisation theory. The realisation step of filter design has to be adapted to the different sensitivity requirements. However the approximation step, performed on the NTF $H(z)$ rather than the loop filter $L(z)$, is still classical. The NTF can be specified with rectangular acceptable / unacceptable regions on the magnitude versus frequency plot in
Choosing the NTF is just a filter design problem. Note that the excluded region (shaded) is on one side only — the Bode integral provides the other constraint. A moderate in-band NTF specification (−40 dB) is again used for clarity.

Practical NTF calculation

It is worth describing how an arbitrary NTF and loop filter are calculated. This is also discussed in (Norsworthy et al. 1997). A filter prototype \( A(z) \) with the desired NTF shape (but probably a gain error, eg. out-of-band gain = 1) is found. For example, a high-pass inverse-Chebyshev \( A(z) \) will result in a conventional low-pass modulator with stopband zeroes. (Experience shows that an inverse-Chebyshev prototype is usually a good choice; an Elliptic prototype wastes less transition bandwidth but usually requires lower out-of-band NTF for acceptable stability.) The NTF has first sample of one, which is satisfied by choosing \( H(z) = A(z)/A(\infty) \). The loop filter \( L(z) \) is calculated from \( H(z) = 1/(1 + L(z)) \).
In practice this is done by equating numerator and denominator polynomials:

\[ H(z) = \frac{1}{1 + L(z)} \]
\[ L(z) = \frac{1 - H(z)}{H(z)} \]
\[ \frac{N_L(z)}{D_L(z)} = \frac{1 - \frac{N_H(z)}{D_H(z)}}{\frac{N_H(z)}{D_H(z)}} = \frac{D_H(z) - N_H(z)}{N_H(z)} \]

The Matlab commands are shown below. Here the filter prototype is a 5th order inverse-Chebyshev with -100dB stopband. The “0.016” is the useful bandwidth (OSR = 1/0.016 = 63) and sets the out-of-band gain. It is adjusted to get \( \|H(e^{j\theta})\|_\infty = 1.6 \) or whatever is desired for stability. The resulting NTF is about +4dB (1.6) out-of-band and -96dB (-100dB + 4dB) over the useful bandwidth.

```
% Prototype
[numa,dena]=cheby2(5,100,0.016,'high')
% Scale to get h0 = 1 in NTF H(z) = h0 + h1 z^-1 + h2 z^-2 + ... 
numh = numa / numa(1)
denh = dena 
% Check out-of-band NTF visually 
freqz(numh,denh)
% Loop filter L(z)
numl = denh - numh
denl = numh 
```

The main point of including this is to show that it is straightforward to convert an arbitrary filter prototype to an NTF and loop filter transfer function.

**Bitstream information capacity interpretation**

A single-loop sigma-delta modulator with a single-bit quantiser which provides 16 bit resolution in-band (98dB SNR) must have an OSR of at least 16. Otherwise the output bitstream would not have sufficient information capacity to represent the input signal. This
sounds banal, but thinking about it can give an indication of whether existing modulators can be significantly improved by yet to be discovered techniques.

The OSR of the 16-bit example above is 63. Therefore, about 1/4 of the bitstream information capacity represents the input signal. What does the other 3/4 represent?

Some bits represent the input signal in the transition band. (In many applications a pre-filter will make this close to zero, but unless the modulator has a peculiar structure which overloads for very small out-of-band inputs, it still consumes bits to represent it.)

This can be estimated by integrating the SNR in bits versus frequency. This is logarithmic in magnitude and linear in frequency like the Bode integral (4.6) at high SNR. Other authors (Gerzon & Craven 1989) (Nawrocki et al 1999) have therefore given the Bode integral an information capacity interpretation. This seems reasonable, although it should be noted that the two expressions behave differently at low SNR. Integrating over the region from \( \theta = 0.016\pi \) (the passband edge, NTF = -96dB) to \( \theta = 0.08\pi \) (where the NTF is around 0dB) indicates that around 30% of the output bitstream capacity represents the transition band. This accounts for 25% (in-band) + 30% (transition) = 55% of the bitstream capacity.

Presumably the other 45% is related to the stopband where the NTF is close to 1.6. However, it is difficult to calculate exactly how much information is present, as small differences between the NTF and SNR lead to large differences in the calculated information content.

The “wasted” 30% transition bandwidth can reduced by using a higher-order loop filter (at the expense of complexity and increased power consumption). How about the 45% suspected to be representing the stopband? If a sigma-delta modulator was found which did not “waste” this bandwidth, and the SNR agreed with the NTF, the modulator would have an out-of-band NTF of a little over two. (Increased out-of-band NTF corresponds to increased useful bandwidth.) This is a useful link between the bitstream information capacity and the Bode integral – information capacity seems to guarantee that an NTF which is too large out-of-band cannot result in a stable modulator. Practical modulators with high-order loop filters are only stable with out-of-band NTF’s up to about 1.75. This
limit (rather than a little over two) is probably the consequence of the raggedness of the
mapping from the previous state and input to the current state. The stability boundaries
(with respect to maximum input, state or parameters) of existing high-order NTF’s tend to
be complicated fractal shapes (Schreier et al. 1997) (Farrell & Feely 1998) (Schreier &
Yang 1992); presumably if a loop filter could be found which more neatly mapped the
previous state and input to the current state, the maximum out-of-band NTF for stability
would increase.

To sum up this discussion, the bitstream information capacity limitation can either be
considered to be manifested mathematically by the Bode integral (as Nawrocki et al. (1999)
and Gerzon & Craven (1989) tend to do), or to impose a necessary limitation on the
out-of-band NTF for stability. The author tends to think in the second way, perhaps out of
familiarity with the mathematics behind the Bode integral, or perhaps because of a stronger
preference for the linearised analysis from experience largely with multi-bit modulators
where it has worked well.

4.4.3 Derivation of integrals

It is helpful to understand where the Bode integral comes from. Sometimes examining the
derivation of a fundamental limitation offers possibilities by showing a premise which can be
negated by a different architecture or circuit. In this case it is more because it leads on to
an additional restriction involving excess loop delay.

The loop filter is

\[ L(z) = l_1 z^{-1} + l_2 z^{-2} + l_3 z^{-3} + \cdots \]

It must be strictly proper (no \( l_0 \) term) or else there would be a delayless path around the
feedback loop. (The notation here is that the quantiser is instantaneous, and there is no
separate \( z^{-1} \) or gain factor elsewhere in the loop. This avoids having to carry a complicated
loop gain expression through the derivation.) The loop filter \( L(z) \) is related to the NTF
\( H(z) \) by

\[ H(z) = \frac{1}{1 + L(z)}. \]
Rearranging and equating coefficients of $z^{-n}$,

\[(1 + L(z))H(z) = 1\]
\[(1 + l_1 z^{-1} + l_2 z^{-2} + \cdots) (h_0 + h_1 z^{-1} + h_2 z^{-2} + \cdots) = 1\]

\[h_0 + (l_1 h_0 + h_1) z^{-1} + (l_2 h_0 + l_1 h_1 + h_2) z^{-2} + \cdots = 1\]

\[h_0 = 1 \quad (4.7)\]
\[l_1 h_0 + h_1 = 0 \Rightarrow h_1 = -l_1. \quad (4.8)\]

Eq. (4.7) has been mentioned; this is where it comes from. Eq. (4.8) will be used when we consider an excess loop delay.

These coefficient relationships can be converted to integral limitations. There are a number of ways to proceed. Bode used contour integration, but a series expansion is used here. This follows Kolmogorov’s work on linear prediction (Shiryayev 1992, vol. 2 p. 272–280).

Let $\log H(z)$ have the series expansion

\[\log H(z) = B(z) = h_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots,\]

so

\[H(z) = e^{B(z)} = e^{h_0} e^{b_1 z^{-1}} e^{b_2 z^{-2}} \cdots.\]

Using the power series for the exponential function,

\[H(z) = e^{h_0} (1 + b_1 z^{-1} + \frac{b_1^2 z^{-2}}{2} + \cdots) (1 + b_2 z^{-2} + \frac{b_2^2 z^{-4}}{2} + \cdots) \cdots\]
\[= e^{h_0} + b_1 e^{h_0} z^{-1} + (b_2 + \frac{b_1^2}{2}) e^{h_0} z^{-2} + \cdots\]
\[= h_0 + h_1 z^{-1} + h_2 z^{-2} + \cdots\]

Equating coefficients,

\[h_0 = e^{h_0} \Rightarrow h_0 = \log h_0\]
\[h_1 = b_1 e^{h_0} \Rightarrow b_1 = \frac{h_1}{h_0}.\]
Around the unit circle \( z = e^{j\theta} \),

\[
\log H(e^{j\theta}) = b_0 + b_1 e^{-j\theta} + \cdots
\]

\[
= \log h_0 + \frac{h_1}{h_0} e^{-j\theta} + \cdots
\]

Now

\[
\log |H(e^{j\theta})|^2 = \log H(e^{-j\theta}^* H(e^{j\theta})
\]

\[
= \log H(e^{-j\theta}) + \log H(e^{j\theta})
\]

\[
= \cdots + \frac{h_1}{h_0} e^{-j\theta} + 2 \log h_0 + \frac{h_1}{h_0} e^{j\theta} + \cdots
\]

Equating the Fourier series coefficients,

\[
\int_{-\pi}^{\pi} \log |H(e^{j\theta})|^2 d\theta = 4\pi \log h_0 \quad (4.9)
\]

\[
\int_{-\pi}^{\pi} e^{j\theta} \log |H(e^{j\theta})|^2 d\theta = \frac{2\pi h_1}{h_0} \quad (4.10)
\]

\[
\int_{-\pi}^{\pi} e^{-j\theta} \log |H(e^{j\theta})|^2 d\theta = \frac{2\pi h_1}{h_0} \quad (4.11)
\]

Adding the last two gives

\[
\int_{-\pi}^{\pi} \cos \theta \log |H(e^{j\theta})|^2 d\theta = \frac{2\pi h_1}{h_0} \quad (4.12)
\]

These relationships look too good to be true for arbitrary \( H(z) \), and indeed this is the case. They rely on the series for \( B(z) \) converging on the unit circle \( z = e^{j\theta} \). This will occur if \( B(z) = \log H(z) \) is analytic for \( |z^{-1}| < 1 \). (Remember that a power series in a complex variable converges out to the radius where it first hits a pole. The power series here are in \( z^{-1} \), not \( z \).) For \( B(z) \) to be analytic in \( |z^{-1}| < 1 \), \( H(z) \) must have no poles or zeroes in \( |z^{-1}| < 1 \) (outside the unit circle). The condition for the integrals above to hold is therefore that \( H(z) \) has no poles or zeroes outside the unit circle.

Let us apply these results to an \( H(z) \) which is a noise transfer function. If the sigma-delta modulator is closed-loop stable, as any practical modulator would be, \( H(z) \) has no poles outside the unit circle. For \( H(z) \) to have no zeroes outside the unit circle, it has been shown previously that the loop filter \( L(z) \) must be stable. (Recall the discussion about
chaotic modulators.) If these conditions hold, substituting (4.7) and (4.8) into (4.9) and (4.12) and using \( \log |H(e^{j\theta})|^2 = 2 \log |H(e^{j\theta})| \) results in

\[
\int_{-\pi}^{\pi} \log |H(e^{j\theta})| \, d\theta = 0 \tag{4.13}
\]

\[
\int_{-\pi}^{\pi} \cos \theta \log |H(e^{j\theta})| \, d\theta = -\pi l_1. \tag{4.14}
\]

Eq. (4.13) is the Bode integral (4.6).

### 4.4.4 Effect of excess delay

Eq. (4.14) is a new integral. It will be interpreted in this subsection. Recall that \( l_1 \) is the first sample of the impulse response of the loop filter \( L(z) = l_1 z^{-1} + l_2 z^{-2} + \cdots \). Also, in the notation used any delay around the modulator loop is included in the loop filter. The quantiser is assumed to be instantaneous; if it has delay then that is included in \( L(z) \).

At high clock rates, it would often be convenient to allow more than one cycle of latency in the loop filter and quantiser combination. The probability of comparator metastability is greatly reduced by passing the comparator output through a few latches. In continuous-time modulators, DAC pulses can be made cleaner. Latency was a significant design limitation for the sigma-delta ADC presented in the next chapter.

If the total latency around the modulator loop is two samples rather than one sample, then the loop filter (including any quantiser delay) is \( L(z) = l_2 z^{-2} + l_3 z^{-3} + \cdots \), i.e. \( l_1 = 0 \). Therefore (4.14) becomes

\[
\int_{-\pi}^{\pi} \cos \theta \log |H(e^{j\theta})| \, d\theta = 0. \tag{4.15}
\]

The question is then whether this integral is satisfied by good conventional NTF’s, or at least something reasonably close. For low-pass modulators, it usually is not. The log of the NTF is very negative at low frequencies. \( \cos \theta \) is close to one there. The log of the NTF is positive at frequencies closer to Nyquist, where \( \cos \theta \) is negative. Both of these result in a negative integrand in (4.15), so it will not integrate to zero. Conventional low-pass NTF’s
like the one in fig. 4.3 therefore cannot be implemented with an excess sample of latency in the loop.

For bandpass modulators with a centre frequency of $f_s/4$, the situation is better. ($f_s$ is the clock frequency. $f_s/4$ is $\theta = \pi/2$.) If the NTF is symmetric around $f_s/4$, (4.15) will be satisfied. The loop filter in this case is of the form $L(z) = l_2z^{-2} + l_4z^{-4} + l_6z^{-6} + \cdots$. Some asymmetric NTF’s are also possible using nonzero $l_3$ etc.

Are there low-pass modulator NTF’s with excess loop delay (no $l_1$ term) and close to optimum performance? To satisfy both (4.15) and (4.13), a little thought shows that the NTF needs to be more positive at frequencies just above the passband, and less positive or negative at high frequencies, than if just (4.13) is satisfied. This uneven out-of-band gain will push up $||H(e^{i\theta})||_\infty$ for a given useful bandwidth, almost certainly degrading stability. Alternatively, for the same stability the useful bandwidth will decrease.

There does not appear to be any NTF which maintains $||H(e^{i\theta})||_\infty$ and satisfies (4.13) and (4.15) without “wasting” about half of the bandwidth. The best way to satisfy (4.15) is for the NTF to be equal to $||H(e^{i\theta})||_\infty$ for all out-of-band frequencies except for a deep notch right at the Nyquist frequency ($\theta = \pi$). This deep notch must have nearly the same area as that of the useful bandwidth at low frequency. Then to satisfy (4.13) the useful bandwidth must be halved compared to the case with no excess delay.

**Continuous-time loop filters**

The desire to use excess loop delay is particularly strong for modulators with continuous-time loop filters and very high clock rates. It is therefore worth explaining how these limitations relate to continuous-time loop filters.

An equivalent discrete-time loop gain $L(z)$ can be calculated by evaluating the samples of the DAC pulse filtered by the continuous-time loop filter fed back to the quantiser at successive sampling instants (Cherry & Snelgrove 19996). If the overall latency around the modulator loop is more than one clock period, then the first sample $l_1$ is necessarily zero. The NTF will therefore be limited by (4.15). This has the same implications as with discrete-time loop filters, namely that either stability or bandwidth suffers.
4.4.5 Excess delay with unstable loop filters

Eq. (4.14) and (4.13) assume that the loop filter is stable. There is therefore a possibility that an unstable loop filter (chaotic modulator) might help. Recall that an unstable loop filter leads to a non-minimum-phase NTF.

To analyse this case, the NTF is factorised into a minimum-phase part $H_{mp}(z)$ and an allpass part $H_{ap}(z)$:

$$H(z) = H_{mp}(z)H_{ap}(z).$$

This approach is standard – mathematicians refer to the minimum-phase part as “outer” and the allpass part as “inner”. The minimum-phase part satisfies (4.13) and (4.14). The allpass part can be factorised into a product of allpass sections

$$H_{ap}(z) = \prod_i H_{ap(i)}(z),$$

where

$$H_{ap(i)}(z) = \frac{k_i z^{-1} - 1}{k_i - z^{-1}}.\quad (4.17)$$

$H_{ap}(z)$ is called a Blaschke product in mathematics. If the $k_i$ are complex, they occur in conjugate pairs. The allpass terms can be integrated using contour integration, giving

$$\int_{-\pi}^{\pi} \log |H_{ap(i)}(e^{j\theta})| \, d\theta = \pi \log |k_i|\quad (4.18)$$

$$\int_{-\pi}^{\pi} \cos \theta \, \log |H_{ap(i)}(e^{j\theta})| \, d\theta = \frac{\pi (|k_i|^2 - 1)}{2k_i}\quad (4.19)$$

Eq. (4.18) is zero or positive. This is where the inequality in (4.6) comes from. The integral contributions from the allpass factors (4.17) add to (4.13) and (4.14), as taking $\log |H(z)|$ converts the products (4.16) and (4.17) into sums.
\[ H(z) = H_{mp}(z) \prod_{i} H_{ap(i)}(z) \]  
\[ \log |H(z)| = \log |H_{mp}(z)| + \sum_{i} \log |H_{ap(i)}(z)| \]

\[ \int_{-\pi}^{\pi} \log |H(e^{j\theta})| \, d\theta = \sum_{i} \pi \log |k_i| \]

\[ \int_{-\pi}^{\pi} \cos \theta \log |H(e^{j\theta})| \, d\theta = -\pi l_1 + \sum_{i} \frac{\pi(|k_i|^2 - 1)}{2k_i} \]  

A positive (4.22) is detrimental as the NTF has to be larger out of band for the same in-band noise reduction. However, this can allow (4.23) to be negative despite the excess loop delay \( l_1 = 0 \). This allows a low-pass modulator with excess loop delay to have lower peak out-of-band NTF (better stability) or lower in-band quantisation noise. The question is therefore whether the benefits from the (possibly) negative (4.23) outweigh the detrimental effects of the positive (4.22).

For \( k_i \) close to \(-1\), (4.22) increases by the same amount amount that (4.23) decreases. The net effect of this is that no more “noise reduction area” is available to improve in-band NTF or stability. Positive or complex \( k_i \) only make matters worse. However, for large negative real \( k_i \) (4.23) decreases faster than (4.22) increases. An unstable loop filter could therefore be beneficial. Most benefit is derived from a single allpass section with large negative \( k_1 \) rather than multiple allpass terms. (This benefit is only to recover the losses caused by excess loop delay; without excess loop delay an unstable loop filter is always detrimental as far as NTF is concerned.) To use a large negative \( k_1 \), the modulator must be stable with sizable out-of-band NTF, as (4.22) is significantly positive. The situation looks more hopeful for multiple-bit quantisers which are stable with greater out-of-band NTF.

**Sample non-minimum-phase NTF**

The easiest way to experiment with such NTF’s is to take an existing minimum-phase NTF \( H_{mp}(z) \), and multiply it by an allpass factor. (In practice, this may be slightly sub-optimal as it raises the degree of the NTF by one without narrowing the NTF transition band.) A raw allpass factor will not leave the \( h_0 = 1 \) coefficient of the NTF unchanged, so it is
necessary to multiply the allpass factor by \(-\bar{k}_1\):

\[
H(z) = -\bar{k}_1 \left( \frac{k_1 z^{-1} - 1}{k_1 - z^{-1}} \right) H_{mp}(z) = \frac{1 - k_1 z^{-1}}{1 - \frac{k_1}{k_1} z^{-1}} H_{mp}(z).
\]

This multiplication by \(-\bar{k}_1\) is how the degradation due to the excess loop delay manifests itself. The NTF is multiplied by \(|k_1|\) at all frequencies, satisfying (4.22). This is primarily a problem out-of-band as it degrades stability. To compensate, a \(H_{mp}\) with smaller \(||H_{mp}(e^{j\theta})||_\infty\) must be chosen. The key question is how much smaller \(||H_{mp}(e^{j\theta})||_\infty\) is for \(l_1 = 0\).

This was tried for two cases; \(||H(e^{j\theta})||_\infty = 1.6\) as might be used for a modulator with single-bit quantiser, and \(||H(e^{j\theta})||_\infty = 5\) as might be used with a four-bit quantiser. A fifth order low-pass modulator with spaced zeroes and \(-96\) dB passband was used. The results were:

\[
\begin{array}{ccc}
||H(e^{j\theta})||_\infty & 1.6 & 5 \\
OSR \text{ with no excess delay} & 63 & 17 \\
||H_{mp}(e^{j\theta})||_\infty & 1.26 & 2.37 \\
k_1 & -1.26 & -2.15 \\
OSR \text{ with excess delay} & 125 & 33
\end{array}
\]

The loss of bandwidth is very close to 50% in both cases. This is also the approximate bandwidth loss without unstable zeroes. Unless a chaotic modulator was desired to decrease tones, an unstable loop filter is of no benefit.

### 4.5 CONNECTIONS WITH OTHER THEORY

The Bode and excess delay integrals may appear to be an obscure corner of circuit theory. However, they have received the attention of a surprising number of notable authors. This section presents two links with other theory. The first is with linear prediction. The author has found it to be conceptually useful, particularly when considering excess loop delay. The second is a discussion of how the theory in this chapter relates to other system theory, such
as the state-space theory in chapter 2. In this case, the links are tantalising but rather hard to pin down.

4.5.1 Linear predictor interpretation

There is another interpretation of the integral limitations which particularly explains why excess delay is detrimental. This is to consider the modulator as a predictive coder. This approach is also useful because linear prediction has been studied by leading mathematicians such as Kolmogorov, Wiener and Szegö, and is therefore well understood.

Older papers such as (Candy 1985) and (Tewksbury & Hallock 1978) consider predictive coding. Such works tend to distinguish between predictive coding and noise shaping modulators, but Tewksbury & Hallock (1978) clearly make the point discussed below that the two are interchangeable by using a pre-emphasis (input) filter. The links are roughly that different modulator structures have similar noise transfer functions but different signal transfer functions. Predictive coding is presumably mentioned less today because the sigma-delta modulator structure has been found to be superior in practice (it is less sensitive to loop-filter gain errors and the like).

The easiest way to link modern sigma-delta modulator structures to predictive coding is with a loop transformation (2A, fig. 4.1). The loop filtering function is split between an input filter and a feedback filter. The input filter shapes the signal input so that it is much larger in-band. The feedback loop then has to predict this weighted input to keep the input to the quantiser small. The overall resolution of the converter is the quantiser resolution multiplied by how accurately the signal is predicted. This is because the prediction is subtracted from the input, leaving only the residual to quantise. (There is another form of the predictive coding modulator in which the prediction is made from the reconstructed signal, not the error signal. This can also be obtained by loop transformation.)

Note that this loop transformation leads to difficulties with unstable loop filters. The case of loop filters with poles on the unit circle (including at DC) would make mathematicians nervous, but engineers would handle it by saying that the integrators or resonators have a very high but finite gain so the poles are just inside the unit circle. This
case was of practical importance for delta modulators, and was handled by including a small integrator “leak” to make the input filter stable (Wooley 2000).

Theoretical limitations on how accurately a coloured noise input can be predicted are well known (Grenander & Szegö 1958) (Scharf 1991). The usual assumptions that the coloured noise is stationary and stochastic are made. The optimum one-step-ahead predictor of a coloured noise \( F(z) \) with power spectral density \( |F(e^{j\theta})|^2 \) has a prediction error (Grenander & Szegö, p. 181)

\[
E(|x_i^t - x_i|^2) = \exp \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |F(e^{j\theta})|^2 d\theta \right)
\]  (4.24)

where \( x_i^t \) is the prediction of \( x_i \) from previous inputs \( x_{i-1}, x_{i-2}, \cdots \) and \( E() \) is expectation. The optimum predictor is the one which whitens the residual error (Bode & Shannon 1950).

The logarithmic integral suggests links with (4.13), and indeed the two are closely related. For the linearised model 2B of the transformed modulator 2A in fig. 4.1, if the input is \( R(z) \) then the signal being predicted is \( F(z) = L(z)R(z) \). From 2B in fig. 4.1, the prediction error is

\[
D(z) = \frac{L(z)}{1 + L(z)}R(z).
\]

The optimum predictor whitens the prediction error, \( D(z) = c \) for a constant \( c \). This results in

\[
\frac{L(z)}{1 + L(z)}R(z) = c
\]

\[
R(z) = c \frac{1 + L(z)}{L(z)}
\]

\[
F(z) = L(z)R(z)
\]

\[
= c(1 + L(z)).
\]

Eq. (4.24) then becomes

\[
c^2 = \exp \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |c(1 + L(e^{j\theta}))|^2 d\theta \right)
\]

\[
2 \log c = 2 \log c + \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |1 + L(e^{j\theta})|^2 d\theta
\]

\[
0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |H(e^{j\theta})|^2 d\theta
\]
using the definition of the NTF, \( H(z) = 1/(1 + L(z)) \).

This is an unnecessarily complicated way of deriving (4.13). The derivation of the NTF integrals in this chapter is actually an intermediate “spectral factorisation” step in the proof of (4.24) (Scharf 1991). The linear prediction interpretation is most useful conceptually rather than for calculation.

The Bode integral (4.13) is thus related to one-step-ahead prediction. With a sample of excess delay in the modulator loop, the problem becomes two-step-ahead prediction. The multi-step prediction error was calculated by Kolmogorov (Shiryaev 1992, vol. 2, p. 274) (and likely also by Wold or Wiener).

The main benefit in relating excess delay to two-step-ahead prediction is conceptual. It becomes a lot more reasonable to expect that excess delay cannot be fully compensated for. The case where one sample of excess delay is not detrimental, namely \( f_s/4 \) bandpass modulators, is instructive. A sample of excess delay is not detrimental because with \( f_s/4 \) centre frequency, the samples go \( I, Q, -I, -Q, \cdots \) and the in-phase signal \( I \) is no use in predicting the following quadrature sample \( Q \). By contrast, for low-pass inputs the previous sample is usually the most important for predicting the current sample. Predictive coders will therefore perform more poorly (require larger quantisation steps and hence have lower SNR) without it. Degradation of the performance of sigma-delta modulators then follows using the loop transformation described above.

Three-step-ahead prediction could be considered, but the mathematics becomes complicated and it is rare to need so much latency. It is also possible to apply the results for prediction of continuous-time processes (Grenander & Szegö 1958, p. 196) to sigma-delta modulators with continuous-time loop filters. However, the results do not seem to have a form particularly conducive to understanding or useful for design.

### 4.5.2 Links with other system theory

The work in this chapter shares many features with the state-space theory in chapter 2. In particular:
• Quantisation noise is analysed as an additive noise in the same way that thermal noise (and perhaps distortion) is in the state-space theory.

• The noise transfer function $H(z)$ in this chapter plays the same role as the noise gain $G(s)$ in the state-space theory.

• Fundamental limitations in both theories follow from requiring the transfer function to be causal. Recall that the Hankel operator was a map from past inputs to future outputs, and that the Bode integral followed from the causality of the noise transfer function.

It is peculiar then that the two theories have such a different mathematical form. Perhaps we are just using different notation for the same thing?

The author has tried to reconcile the two theories but has made little progress. They are not identical – most notably, the quantisation noise is only added at one point, whereas the state-space theory had a noise source adding to each state. This may be why the Bode integral is a weaker logarithmic restriction on the dynamic range. A future theory which could deal with noise added to only some of the states would surely be useful.

It should be noted that others have almost certainly also tried to find links between Bode integrals and state-space realisation. The Bode integral and the frequency-weighted balanced realisation are generally considered to be amongst the most important results in robust control theory. One of the finest recent textbooks, (Zhou 1996), has the two theories in successive chapters. No link is made between them.

Logarithmic integrals also occur in broadband impedance matching theory (Bode 1945) (Fano 1950). Interestingly, Helton (1981) has applied Hankel operators to this theory. Helton connects the theories by stating that the logarithmic integral uses one basis for the Hilbert spaces involved, and the Hankel operator theory uses another. The logarithmic integral basis is notable for giving limitations in a finite number of dimensions, but is difficult to use for automatic computations.
4.6 DESIGN IMPLICATIONS OF THE INTEGRAL LIMITATIONS

For standard low-pass modulators with one-bit quantisers, most of the design limitations are well enough known that the theory developed here does not predict any new avenues to explore. The most useful feature of the integral limitations is perhaps the negative result that resolution for a given OSR can only be improved by increasing the out-of-band noise gain, or by sharpening the loop filter response.

It is worth mentioning one other point which may have occurred to the reader. It is known that spreading the NTF zeroes widens the bandwidth. The NTF zeroes are still on the unit circle. Aren’t the deep notches in the NTF a waste of noise reduction area – wouldn’t it be better for the NTF to be constant across the passband? This is the case, but a flatter in-band NTF would require a higher-order loop filter for a given transition bandwidth so there is little advantage.

4.6.1 Multiple-loop modulators

The theory so far concentrates on single-loop modulators. However, the best sigma-delta modulators are often cascades of low-order loops (the MASH architecture) (Norsworthy et al. 1997). For multiple-loop modulators, performance is usually limited by a combination of the NTF of the first loop and the degree of matching between loops (fig. 4.5). By applying the integral limitations to the first loop, useful conclusions can still be drawn.

Suppose that the first loop of a multiple-loop modulator has −36dB or less quantisation noise in-band. The second and subsequent stages cancel this noise. If the second stage DAC has a 0.1% (−60dB) gain error due to capacitor mismatch and the like, the lowest quantisation noise which is possible is around −36dB − 60dB = −96dB. To satisfy the Bode integral (4.6) with an out-of-band gain of \(\|H(e^{j\theta})\|_\infty = 1.6 \ (+4\ dB)\), the OSR must be

\[
-36\ dB \left( \frac{1}{\text{OSR}} \right) + 4\ dB \left( 1 - \frac{1}{\text{OSR}} \right) \geq 0\ dB
\]

\[
\text{OSR} \geq 10
\]

The more general result is that the SNR achieved by a multiple-loop modulator is limited to that achieved by a single-loop modulator in accordance with the Bode integral, plus the
Figure 4.5: NTF’s for the first loop of a multiple-loop (MASH) modulator. The fundamental limit to performance is usually the NTF (solid curves), of the first loop, multiplied by the mismatch error (here 0.1% or −60 dB) between the first and second loop (dotted curves). Note that the second-order loop-filter is not much worse than the fifth order loop-filter.
degree of matching.

This analysis predicts that the best performance would be to use a first loop with high-order loop filter and spaced NTF zeroes to achieve the \(-36\) dB NTF over as wide a band as possible. Practical design deviates from this for two reasons. Firstly, only a second-order loop filter is usually used in the first loop as it is absolutely stable and simpler. It is also stable with higher out-of-band gain that \(\|H(e^{j\theta})\|_\infty = 1.6\) (Schreier & Yang 1992), which at low OSR almost compensates for the larger wasted transition bandwidth. (The transition band is a lot narrower if only \(-36\) dB NTF is required rather than \(-96\) dB NTF.)

Secondly, spaced NTF zeroes seem to cause larger tones (Thoss et al. 2000) (Vleugels 2000). It may be that this would be less of a problem if a higher-order loop filter was used.

### 4.6.2 Multiple-bit quantisers

For multiple-bit quantisers, the theory seems more useful. The main prediction is that large out-of-band NTF’s are the way to get high bandwidth at very low OSR (4–10). (Ju & Vallancourt 1992) have shown that \(\|H(e^{j\theta})\|_\infty = 5\) (+14 dB) is stable with a four-bit quantiser. The stability is for inputs up to 0.8 of the maximum DAC output so presumably it is robust. To get 80 dB SNR with a four-bit quantiser, an NTF around \(-60\) dB is required. A fifth-order loop filter with spaced zeroes and +14 dB out-of-band gain can an achieve this at OSR=8.

Conventional wisdom (Norsworthy et al. 1997, p. 244) is that the SNR benefit of multi-bit quantisers comes from the 6dB / bit reduction in raw quantisation noise. It is widely known that a multi-bit quantiser gives a more stable modulator, but this is seen only as a incidental bonus. At most, an NTF of the form \(H(z) = (1 - z^{-1})^n\) with \(n\) greater than two might be used.

The integral limitations make quite a different prediction. To maximise bandwidth, out-of-band NTF must be aggressively maximised, constrained of course by stability for a reasonable maximum input. The NTF zeroes should be spread across the passband. Fortunately, multiple-bit modulators usually have few problems with tones and their stability correlates well with the linear NTF-based model.
4.6.3 Compensation for excess loop delay

The restrictions resulting from excess loop delay have been calculated in detail in previous sections. The aim here is to discuss the practical effects, particularly whether a performance degradation is inevitable or whether it can be compensated for by changing the loop filter. The theoretical results will also be checked against simulations in the literature. As the sections on excess loop delay were detailed and mathematical, a reminder of the results is in order:

- A modulator with one sample of excess delay around the modulator loop \((l_1 = 0\) in \(L(z) = l_1z^{-1} + l_2z^{-2} + l_3z^{-3} \cdots\)) and no loop filter poles outside the unit circle obeys (4.13) and (4.15). To satisfy both of these, the NTF of a low-pass modulator is inevitably distorted in such a way that either the peak out-of-band NTF increases or the useful bandwidth decreases. The optimum NTF with excess loop delay is one which has a large wide notch near the Nyquist frequency to satisfy (4.15) with minimum loss of “noise reduction area” in (4.13). This still halves the useful bandwidth.

- A modulator with or without loop filter poles outside the unit circle obeys (4.22) and (4.23). Loop filter poles outside the unit circle give a chaotic modulator (Schreier 1994). The limitation resulting from (4.22) and (4.23) is difficult to assess exactly, but sample NTF calculations show that there appears to be negligible benefit in using an unstable loop filter.

- For continuous-time loop filters, there should be no loss of bandwidth provided that some feedback gets around the loop in less than one clock period.

- Interpreting the sigma-delta modulator as a predictive coder, it seems reasonable that performance should necessarily be degraded by excess delay, as the modulator has to predict its input further into the future. For loop filter poles outside the unit circle, this interpretation is not really valid as the process being predicted is unstable.

- Bandpass modulators with \(f_s/4\) centre frequency can achieve the same performance with a sample of excess delay.
Considering these theoretical results, it does not seem possible to fully compensate for excess delay in low-pass modulators. It is reasonable to expect that a loop filter optimised for use with excess delay will perform better than one optimised for no excess delay. However, the theory shows that the best sigma-delta modulator with an excess delay of one sample will still only have about half the useful bandwidth of the best one without excess delay.

Reading papers such as (Cherry & Snelgrove 1999b) and (Yahia, Benabes & Kielbasu 2001), it is easy to come away with the impression that excess delay can be compensated for without bandwidth reduction. Which is wrong, the theory here or the literature? Closely examining the results and discussion in (Cherry & Snelgrove 1999b), this author cannot see anything which contradicts the theory presented here. Cherry & Snelgrove (1999b) present simulations which show the following:

- Excess delay can be fully compensated for in bandpass modulators with \( f_s/4 \) centre frequency.
- Low-pass modulators can be stable with excess delay, but it is only achieved at the expense of bandwidth or resolution by using lower \( \|H(e^{j\theta})\|_\infty \).
- A return-to-zero DAC works well with less than \( 1/2 \) cycle of delay. As Cherry & Snelgrove (1999b) point out, this allows an \( l_1 \) to be realised.

All of these results agree with the analysis here. It is only the conclusion “excess delay can be rendered effectively a non-problem in high-speed CT \( \Delta \Sigma \)M’s” (Cherry & Snelgrove 1999b, p. 388) which is contentious. Certainly a loop filter which has been optimised for excess delay (by making the NTF droop at high frequencies as discussed above) is better than one optimised without loop delay, but there is still an inherent loss of about 50% of the bandwidth for a given resolution. To compensate for excess delay, (Cherry & Snelgrove 1999b) recommend using an NTF with \( \|H(e^{j\theta})\|_\infty = 1.3 \) before the excess delay (of one sample) is added. Matlab NTF calculations show that this requires OSR=112 for \(-96\,\text{dB} \) NTF with a fifth order modulator with spaced zeroes, as compared to OSR=63 with no excess delay and \( \|H(e^{j\theta})\|_\infty = 1.6 \).
There could conceivably be nonlinear effects which degrade the NTF but preserve the SNR. This is not the case - the simulations in (Cherry & Snellgrove 1999b, fig. 14) show a clear difference in SNR between an NTF with $\|H(e^{j\theta})\|_\infty = 1.6$ and no excess delay, and an NTF of $\|H(e^{j\theta})\|_\infty = 1.3$ with excess delay. For a 5th order modulator with spaced zeroes and an OSR of 64, the dynamic range is 19.2 bits with $\|H(e^{j\theta})\|_\infty = 1.6$ and no excess delay, and 15 bits with $\|H(e^{j\theta})\|_\infty = 1.3$ (before excess delay) and one sample of excess delay. These results are for continuous-time modulators, but given the NTF equivalence there should be little difference for discrete-time modulators.

Similarly, (Yahia et al. 2001, p. 344) claims that “Theoretically, high loop delay can be achieved and non-ideal DAC can be used without effect on the performance of the modulator, if some special feedback schematics are used”. However, in the body of their paper they acknowledge that a feedback term necessary for compensating for quantiser delay (their “location #1”) degrades the NTF and produces a resolution loss.

To end this discussion on a more positive note, it should be acknowledged that different modulator architectures are still useful in combating excess loop delay. The theory shows that for continuous-time low-pass modulators, some form of feedback must get around the loop in one sample to realise the $l_1$ in the loop filter $L(z) = l_1 z^{-1} + l_2 z^{-2} + \cdots$. There are good and bad ways of doing this. The best that the author has seen is to have an extra DAC which feeds directly to the quantiser input rather than through the integrators (Luh, Choma & Draper 2000). This is good because the output of that DAC only needs to be ready after one sample of delay, not after just one half of a sample of delay as would be required with a return-to-zero DAC feeding an integrator. The other DAC’s feeding the integrators can have longer latency. A similar “express” feedback path is used in the ADC in the next chapter.

The other positive outcome which is clarified by the theory is that there is no reason not to operate $f_s/4$ bandpass modulators with a sample of excess delay.
4.7 CONCLUSIONS

The noise transfer function (NTF) is constrained by the Bode integral (4.6). This integral limitation is clearly expressed as a conservation of area on a plot of NTF in decibels versus frequency on a linear scale. The limitation also has interpretations in terms of information capacity of the output bitstream, and in terms of predictive coding. The following conclusions follow from the integral limitation:

- The oversampling ratio (OSR) for a given resolution depends on two factors: how large the out-of-band NTF can be made whilst retaining acceptable stability, and how much “noise reduction area” is wasted in transition regions of intermediate NTF. The traditional curves relating resolution to OSR and loop filter order are not mysterious; they follow simply from these factors.

- Modulators with multiple-bit quantisers necessarily need to be operated with high out-of-band NTF’s to achieve the widest bandwidth for a given resolution. The advantage that they are stable with higher out-of-band NTF is as important or more important than the 6dB reduction in raw quantisation noise with each extra quantiser bit.

Another integral limitation (4.15) shows that a sample of excess delay around the modulator loop roughly halves the useful bandwidth achievable for low-pass modulators. “Achievable” includes attempts to compensate for excess loop delay. The result is derived in discrete-time, but it applies to the important case of modulators with continuous-time loop filters if no DAC pulse returns to the quantiser in time for the next sample. Bandpass modulators with $f_s/4$ centre frequency need not lose bandwidth. Simulations in the literature are consistent with this.
5 A 520MHZ SIGMA-DELTA ADC WITH A FIR DAC LOOP FILTER

5.1 INTRODUCTION

This chapter describes a high-speed, low-resolution sigma-delta ADC which was fabricated in a 0.25 μm process. The aim was to build a converter with bandwidth beyond that practical with switched-capacitor circuits. The ADC achieves 60 MHz signal bandwidth, albeit at only 40 dB signal to noise and distortion ratio. The ADC uses a novel finite-impulse-response (FIR) digital to analog converter (DAC) structure in place of a conventional loop filter.

This section introduces the FIR DAC and the reasons for using it. Section 5.2 discusses choice of noise transfer function. The theory in the previous chapter suggested using a multi-bit quantiser and high out-of-band NTF; the converter presented uses a four-bit quantiser and +17 dB peak out-of-band NTF. Section 5.3 describes the architecture at the block level, and section 5.4 describes it at the circuit level. Section 5.5 describes the testing procedure and test printed circuit board. This section is included as testing high-performance ADC’s is non-trivial. The measured results are given in section 5.6. The final parts of the chapter describe possible modifications to the architecture, and compare it with other approaches such as pipelined converters.

5.1.1 Argument for FIR DAC loop filters

The previous chapter shows that a sigma-delta modulator can achieve high resolution at low oversampling ratio (OSR) if the noise transfer function (NTF) is high out-of-band. A multi-bit quantiser is needed for stability with high out-of-band NTF. (Intuitively, a multi-bit quantiser has less quantisation noise, so it takes a higher out-of-band gain for the quantisation noise to overload it.) Multi-bit quantisers also offer increased resolution even
before the out-of-band gain is increased, simply because the quantisation steps are smaller.

The definition of NTF is identical to that of sensitivity. A high NTF implies that a small gain or phase error can make the modulator unstable. To realise the advantages of high out-of-band NTF, the loop filter $L(z)$ must be accurately realised.

At clock frequencies below $200$ MHz, switched-capacitor circuits allow an accurate loop filter to be realised. This approach has been heavily studied (Norsworthy et al. 1997) (Candy & Temes 1992), so the aim here was to try to operate at clock frequencies above those possible with today's switched-capacitor circuits. These frequencies are currently the domain of sigma-delta modulators with continuous-time loop filters.

Almost all modulators built with continuous-time loop filters have used one-bit quantisers and hence out-of-band NTF around $1.4$–$1.7$. An argument for one-bit quantisers is that they are inherently linear, so the converter resolution is not limited by transistor matching. However, such converters rarely achieve resolutions beyond the matching level of MOS transistors (often $10$–$12$ bits), so there is no need for one-bit quantisers. If higher resolutions are needed, there is also an opportunity to use dynamic element matching. One-bit quantisers also use less power, but in CMOS the transistors are small enough that $3$-$4$ bit quantisers are quite practical.

A more serious reason why sigma-delta modulators with continuous-time loop filters cannot be operated with out-of-band NTF's of $+15$ dB or so is that the continuous-time loop filters simply are not accurate enough. A gain controlled to $0.5$ dB and a phase controlled to $10^\circ$ is required. It seems unlikely that structures such as Q-enhanced LC resonators (Gao & Snelgrove 1998) or integrators based on simple differential pairs could achieve this level of control. The opamp-RC filter technology in chapter $3$ might be suitable. This filter work was done after the ADC described here was fabricated, so it was not an option at the time. A different modulator structure, using a FIR DAC, was used. The relative merits of FIR DAC and continuous-time loop filters are discussed later in the chapter.
Loop transformation

1. Conventional sigma-delta modulator

2. Transformed "delta" modulator

1A. Continuous-time loop filter

2A. FIR DAC architecture

2B. Architecture with digital filter

Figure 5.1: Wideband sigma-delta ADC architectures
5.1.2 FIR DAC loop filter

The sigma-delta modulator presented in this chapter realises a finite-impulse-response (FIR) transfer function in the DAC. The DAC current fed back is a weighted sum of the last 12 quantiser outputs. The quantiser has four bits, or 16 levels. The DAC's are unary (thermometer-coded), so in total there are $16 \times 12 = 192$ switched current sources in the FIR DAC.

How this can replace the loop filter in a conventional sigma-delta ADC can be seen from a loop transformation (fig. 5.1). In a conventional sigma-delta ADC, the loop filter filters both the input signal and the fed-back signal. In this modulator, these two functions are handled by separate filters – an analog pre-filter and the FIR DAC. The requirements on the analog pre-filter are similar to those of an anti-alias filter for a Nyquist-rate converter. For simplicity, it was omitted from the test chip – an opamp-RC filter like that in chapter 3 would easily meet the requirements. It is important to point out that there is a big advantage in making this split – the analog pre-filter is no longer in the feedback loop, so NTF and stability do not depend on its accuracy. Its phase response can be arbitrary, and its gain only needs to be flat to within a few decibels in the passband and to be sufficiently low out-of-band. The FIR DAC does need an accurate transfer function, but DAC's and digital delays are naturally more accurate.

Another way of looking at this loop transformation is that the converter is a delta modulator (albeit with unusual NTF) rather than a sigma-delta modulator (Wooley 2000).
5.2 CHOICE OF NOISE TRANSFER FUNCTION

The NTF is constrained by the loop filter being of the form

\[ L(z) = l_1 z^{-1} + l_2 z^{-2} + \cdots + l_n z^{-n}. \]

The NTF was numerically optimised in Matlab using the Nelder-Mead simplex algorithm optimiser \textit{fmins}. A variety of FIR lengths \(n\) were tried before settling on \(n = 12\). Higher OSR’s require higher \(n\). A one-norm constraint on the noise gain (Kenney & Carley 1993) (Schreier & Yang 1992) was used, as this offers a guarantee of stability. This was discussed in chapter 4. The in-band NTF was minimised subject to this, for an OSR of around five. (A range of OSR’s were tried; OSR=4–6 was most likely to exceed the state-of-the-art in CMOS ADC’s as >100 Ms/s is difficult for pipelined converters.)

High-order single-loop sigma-delta modulators are usually conditionally stable, and hence require a reset. In early simulations this modulator was no exception. Implementing a reset turned out to be surprisingly difficult. During reset the loop gain has to be reduced by over 30 dB so that the loop gain is below 0 dB in-band. Reset switch capacitance increased the error amplifier settling time. Error amplifier offset voltage and the state of the FIR DAC meant that recovery from reset was somewhat uncertain. It was therefore a pleasant surprise when one choice of FIR taps revealed an absolutely-stable modulator in circuit simulation. What had happened was that the maximum phase lag in this simulation was slightly less, around \(-170^\circ\) (except for frequencies where the loop gain was less than 0 dB) as compared to \(-200^\circ\) in others.

An additional constraint was therefore added to the NTF optimisation to make sure that the phase of \(L(z)\) could be no worse than \(-160^\circ\) at frequencies where the gain exceeded 0 dB. All resulting modulators were absolutely stable, and the performance degradation from the additional constraint was small. (The condition (Bode 1945) that an amplifier is absolutely stable if the phase lag does not exceed 180° is an approximation – see (Vidal 1969). Counter-examples led to the development of rigorous criteria such as the circle and Popov tests for continuous-time, and the Jury-Lee test for discrete-time (Vidal 1969). Nevertheless, Matlab simulations show that the \(< 180^\circ\) phase lag criterion is fairly accurate for the sigma-delta modulators examined.)
ARCHITECTURE DETAILS

The architecture is basically that of 2A in fig. 5.1. A more detailed stage-by-stage diagram is fig. 5.6.

5.3.1 Error amplifier

In-band, the error signal which is fed to the quantiser is about −30 dB relative to the input and FIR DAC signals. This is like any feedback loop - the signal fed back closely matches the input signal. Now the four-bit quantiser operates with signal levels around 1 V, so the error signal must be amplified before being quantised. (It is conceivable that the input and DAC currents, each around 3 mA, could be subtracted, and the difference across a resistor of around 10kΩ fed straight to the quantiser. However, speed and DAC accuracy would greatly suffer from using such a high load impedance.)

The error amplifier gain used was around 25 dB. The error amplifier consists of cascaded differential pairs. As only the error signal is being processed, the error amplifier does not
Figure 5.4: Loop gain of the described ADC. Note that the phase lag does not exceed 180° when the gain is > 0 dB, resulting in an absolutely-stable modulator.
need to operate with the full precision of the ADC. Around 30 dB of dynamic range is sufficient. Therefore, feedback amplifiers are not required. This is vital considering that the error amplifier has to settle in less than one cycle at 520 MHz, i.e. 2 ns.

**Replica reference ladder**

The disadvantage with simple differential pairs is that they do not have accurately known gain. This problem was overcome by feeding the reference voltages for the quantiser through replica differential pairs (fig. 5.6), giving a closely-controlled error amplifier / quantiser overall gain.

**Track and hold**

The input signal is continuous-time, and the signal fed back from the FIR DAC is discrete-time. At some point the input signal must be sampled. This could be done by the quantiser, but the error signal would have a very high $\frac{dv}{dt}$ by that point. This would make both error amplifier and quantiser design very difficult.

A track and hold at the input could also be used. But it is better to sample the error
signal rather than the input signal, as this reduces the required track and hold dynamic range by the in-band loop gain (30 dB). This is a considerable advantage of this architecture over Nyquist ADC’s. The track and hold dynamic range required is only around 30 dB, so simple series CMOS switches and hold capacitors can be used.

The track and hold was eventually placed in the middle of the error amplifier. Some gain (around 8 dB here) before the track and hold decreases its thermal noise contribution. Too much makes the \( \frac{\text{d}v}{\text{d}t} \) too high. Thermal noise is not a limiting factor for this converter, but in the future this advantage over Nyquist ADC’s may prove to be of considerable importance.

5.3.2 Latency and first-tap problems

The greatest design difficulty is that an error signal must be fed back with a latency of only one cycle. This is not a particular drawback of the architecture used here; it is inherent to all conventional single-loop low-pass sigma-delta modulators. We saw in chapter 4 that omitting a \( L_1z^{-1} \) term from the loop gain \( L(z) = L_1z^{-1} + L_2z^{-2} + \cdots \) roughly halves achievable bandwidth.

Getting a signal through an error amplifier, track and hold, quantiser, DAC and subtractor in 2 ns (1/520 MHz) is genuinely difficult. The comparators in the quantiser, and the error amplifier each basically take half a cycle (1 ns) to settle, and that does not leave time for the track and hold. It was therefore necessary to feed the \( L_1z^{-1} \) FIR tap back after the track and hold. This increases the error signal which includes only the \( L_2z^{-2} + \cdots + L_{12}z^{-12} \) terms, requiring a little more linearity in the early stages of the error amplifier. It also leads to gain mismatches between the \( L_1z^{-1} \) term and the other ones. The DC gain errors can be overcome using replica structures, but the dynamic mismatch is a problem.

Overall, the speed of the converter is limited by the error amplifier settling time, particularly for the \( L_1z^{-1} \) FIR tap. At 520 MHz the error amplifier only settles to about 85% of its final value for the first tap, and 90% for the others. The resulting loop gain is therefore

\[
L(z) = 0.85L_1z^{-1} + 0.9L_2z^{-2} + 0.9L_3z^{-3} + \cdots + 0.9L_{12}z^{-12}.
\]
This needs to be taken into account when choosing the tap coefficients $l_1$-$l_{12}$. Typical process variation should not be enough to alter these numbers enough to cause instability (about 10% difference is needed).

5.4 CIRCUIT AND LAYOUT DETAILS

The circuits used for the error amplifier, quantiser, shift register and DAC are conventional. All signals are fully differential to maximise power supply rejection and common-mode rejection. It would be practically impossible to use single-ended signals as the package has significant inductance and digital power supply noise is present. A stage-by-stage diagram of the ADC is shown in fig. 5.6, and a die micrograph is shown in fig. 5.7.

5.4.1 Error amplifier

The stages in the error amplifier are nMOS differential pairs with pMOS diode-connected transistors as loads (fig. 5.6). By making the pMOS devices narrower than the nMOS devices, a gain per stage of around 8 dB is achieved. Diode-connected pMOS transistors give slightly less distortion than resistors as loads, because the nonlinearity in the nMOS differential pair and pMOS loads tends to cancel. The only exception is the final stage which drives the comparators, which is of the same form as those used in the filter in chapter 3. The resistors in its common-mode feedback are replicas of the reference ladder (fig. 5.6). It also has resistive source-degeneration to decrease distortion. Each signal (input, main feedback and first-tap feedback) passes through identical cascades of differential pairs. The track and hold is after the first error amplifier stage. It uses pMOS switches with dummy pMOS devices to cancel charge injection (these may not be necessary). The hold capacitors are the gate capacitance of the second stage – this is sub-optimal with respect to linearity, but the dynamic range required is only 30dB or so and speed is critical.
Figure 5.6: ADC block diagram and typical gain stages. Note the identical parallel signal paths in the error amplifier.
Figure 5.7: ADC die micrograph
Figure 5.8: Comparator and latch (blocks labelled “Comp” and “Latch” in fig. 5.6). (Note that the comparator is preceded by differential pairs, not shown.) Comparator transistors are all 3/0.24, except for 1.44/0.24 in the inverters.

5.4.2 Quantiser

The quantiser is a four-bit flash ADC. As the state of the art is six bits at above 1 Gs/s in a few hundred milliwatts (Choi & Abidi 2001), a quantiser which achieves four bits at 520Ms/s in 30 mW is not particularly difficult. The comparator consists of a pair of differential pairs (to compare a differential reference to a differential input), followed by a differential pair as second-stage preamp, a comparator of the same topology as that in (Kim, Kusayanagi & Abidi 1997) and a differential conventional CMOS latch – see fig. 5.8. Comparators with differential-pair input were preferred to those using switches as they avoid injecting charge onto the reference ladder, and the offset voltage of a differential pair is acceptable for a four bit quantiser.

5.4.3 Shift register and DAC

The shift register uses a latch built in the conventional static CMOS manner from four CMOS switches and four CMOS inverters (fig. 5.8). It has differential input and differential output so that the DAC is driven with a symmetrical signal. Switch transistors are 0.72/0.24 and inverter transistors are 1.44/0.24 (nMOS = pMOS). A single-phase clock is distributed and the complement generated locally to avoid skew between the clock and its
complement. A large clock spine runs along the middle of the converter (fig. 5.7). It is 18μm wide, 1.2mm long, and consists of metal four and five in parallel. The clock buffer drives both the converter and a 128 stage shift register which allows output data to be stored and read out later at a much lower rate. The clock buffer transistors are 3840/0.24 pMOS and 1920/0.24 nMOS.

The clock input from off-chip is differential. This greatly reduces problems with clock jitter from ground bounce. A peculiar circuit based on the author’s work with quadrature generators (Harrison & Weste 1999) was used, although a more conventional ECL to CMOS converter would have worked equally well. There is also a synchronous clock enable which allows the clock to be latched whilst changing over to a lower frequency to clock out the 128 stage test shift register data.

The DAC cell is conventional, with 1.44/0.24 nMOS switches and an uncasced 3/3 nMOS current source transistor. The DAC outputs operate into an nMOS common-gate buffer, so a cascoded current source is not required. (This common-gate buffer is needed to provide a high output resistance at the nodes where the input signal and FIR DAC feedback are summed into a resistor. This resistor matches the reference current input, so any transistor output conductance would cause a gain error.)

The DAC cells operate at currents from 0.5μA to 30μA, depending on which tap (fig. 5.5) they are 1/16 of. The 12 reference biases are generated by a small feedback amplifier which provides the same bias to the DAC cell as it feeds back to the reference input pin. More will be said later about this circuit as it caused problems during testing.

5.5 TEST CIRCUIT BOARD AND TESTING

Testing of an ADC with such a high clock rate is non-trivial. The clock rate was too high for any available logic analyser to store the digital output. Achieving full dynamic range required careful board layout and consideration during integrated circuit design. Useful information on test PCB layout for high-speed ADC’s is provided in ADC manufacturer data sheets, such as (Analog Devices 2000). The test printed circuit board (PCB) is shown
5.5.1 Data and clock inputs

The data and clock inputs are both differential. Baluns consisting of a twisted pair around a ferrite toroid were used to generate the differential signal from a 50 Ω BNC input. A small RC low-pass filter with a cutoff around 50 MHz was placed right next to the signal input pins, to minimise high-frequency noise on the inputs. This was needed because with no input pre-filter on chip, the ADC was sensitive to overload from even small out-of-band signals. The clock power was fairly large so as to minimise jitter due to digital noise and ground bounce. About +5 dBm was the minimum for reasonable operation; +20 dBm would destroy the clock inputs. There were a few bad clock frequencies where the minimum clock input for correct operation was larger and the in-band noise floor was higher, most notably from 550–600 MHz. This suggests some resonances in the clock or power lines.
5.5.2 Digital output

It was originally intended that the output data would be stored in the 128 bit test shift register and clocked out later. This worked, but a 128 point FFT has little frequency resolution. (Only 128 samples were available due to chip area restrictions on the shift register. The 16 levels were stored as thermometer code rather than four bit binary so that bubbles in the quantiser could be seen and so that ground bounce would be linear with output.) It was also possible to use the 16 digital outputs as primitive unary (thermometer-coded) DAC using 16 off-chip 4.7 kΩ chip resistors and a 51 Ω termination. This seemed to give the same SNDR, so in later testing it was used instead. The primitive DAC output was fed directly into a spectrum analyser. For future converters, an on-chip DAC seems natural – it makes testing fast and simple, and as ADC performance lags DAC performance it is not difficult to design the DAC.
5.5.3 Power rails

The chip had analog $V_{DD}$, digital $V_{DD}$, and a shared ground. The logic behind sharing analog and digital ground was that they are connected by less than 1Ω to each other via the substrate anyway, and so may as well share pins. This resulted from a failure to understand that digital $V_{SS}$ is usually kept separate from the substrate. The author presumed that keeping digital $V_{SS}$ separate would cause latch-up problems. This misunderstanding was not revealed by reading several papers on substrate noise, so the author points it out to textbook and expository article writers. Nevertheless, the shared digital ground was not a serious problem as it allowed more pins (four, limited by chip perimeter) to be used for ground. The use of a thermometer-decoded digital output also helped as it meant that ground bounce was linear with the input signal. This means that it should not cause nonlinearity if coupled back into the analog circuitry.

The power rails were bypassed with off-chip 1μF surface-mount ceramic capacitors as close as possible to the IC package. A low-profile IC socket was used to minimise inductance. The circuit is not too sensitive to decoupling as it has about 1nF of on-chip decoupling capacitance. Package inductance was included in simulation. It was originally feared that the on-chip decoupling capacitance would resonate with the package inductance to such a degree that off-chip series resistors of about 0.5Ω in the power rails would be required. However, simulation of the nMOS / pMOS decoupling capacitors used showed that their Q was only about two and there would not be a problem. The analog and digital $V_{DD}$ were fed from the same LM317 regulator (set to 2.5V) via RF chokes and ceramic and tantalum chip decoupling capacitors.

5.5.4 Reference currents

The quantiser resistor ladder reference current and error amplifier and comparator bias currents were supplied by LM334 current sources. The quantiser reference is a differential input so it did not have ground bounce difficulties. The bias inputs have fairly low input resistance and high input capacitance and did not cause problems. (On-chip bypassing using gate capacitance should have been used however.) There is also a gate bias voltage used by
the common-gate stages in the DAC buffer and input amplifier. It had on-chip bypassing as it had oscillated in simulations with package inductance. It caused no further problems.

The pins which caused the most problems were the reference currents which control the FIR tap coefficients. As mentioned earlier, these have small feedback amplifiers in them. These have a bandwidth of around 100MHz. The reference currents were set by series trimpots, with a fixed resistance in series with the trimpot used to measure the reference current. The PCB traces to the trimpots tended to resonate. There are 12 of them, so the probability of one resonating at the clock frequency was high. The feedback amplifiers run at different currents, which leads to them having different gain-bandwidths. They are slightly underdamped, heightening the resonances excited by ground-bounce energy. The problem was severe enough that the first PCB did not work at clock frequencies above 100MHz, and had good and bad clock frequencies. The resonances do not cause nonlinearity, but the change the tap coefficients to such a degree that the sigma-delta modulator oscillates.

The resonances in the FIR reference current lines were overcome on the second PCB by including 1 kΩ series chip resistors close to the IC package. Nevertheless, they may still have had an effect – the degradation of the SNDR as the clock rate was increased from 520MHz to 760MHz (the maximum for correct operation) was uneven. (At frequencies above 520MHz, the taps, particularly the first, also had to be retuned.)

5.6 MEASURED PERFORMANCE

The best performance that might be expected is as follows. An ideal four-bit quantiser has a signal to noise ratio of 26dB. The NTF averages about –25dB over the passband at an OSR of 5. There is a little additional noise reduction from oversampling due to the noise being spread over a wider bandwidth. At an OSR of 5, this is $10 \log_{10} 5 = 7$ dB. To avoid overloading the quantiser from a combination of input signal and quantisation noise, the maximum input is around –6dB. Therefore the overall SNR is $26\text{ dB} + 25\text{ dB} + 7\text{ dB} - 6\text{ dB} = 52\text{ dB}$. 

172
The measured SNDR (signal to noise and distortion ratio) was 40 dB across a 60 MHz bandwidth (an OSR of 4.3) when clocked at 520 MHz. It was not significantly better at higher OSR. The difference between this and ideal is largely due to two factors:

1. The comparators in the quantiser have significant offset. (Comparator transistor size is a trade-off between offset and capacitive load on the error amplifier.) It is estimated that this reduces the SNDR of the quantiser from 26 dB to about 20 dB.

2. Error amplifier distortion reduces the input level for optimum SNDR by about 6 dB. Separate feedback for the first FIR tap worsens the problem as the error signal is larger before the first FIR tap feedback is subtracted. This problem could have been picked up in simulation, but it was difficult to spot as simulations had to run for a day to get even a rough FFT.

Measured performance is shown in fig. 5.11, fig. 5.12 and the table below. A clock frequency of 520 MHz was generally used as the clock jitter tended to vary with frequency and 520 MHz was the highest frequency before it hit a bad patch around 550–600 MHz, above which the converter was running too fast for the error amplifier to settle properly. (The problem was presumed to be clock jitter because the observed effect was an in-band noise floor which depended on clock amplitude.) An SNDR of 30 dB could be obtained with 760 MHz clock and some retuning of taps. The noise floor does not vary with input signal (except at very high input amplitudes). In-band tones do not seem to be a problem – the combination of a multi-bit quantiser and thermal noise presumably breaks them up. Intermodulation distortion (and hence SNDR) does not vary much with signal frequency below 50 MHz. Above 60 MHz, the maximum input falls off quickly with increasing input frequency. (Remember that the required pre-filter was not implemented.)
Figure 5.11: Two-tone measurement (38 MHz, 40 MHz) - 10 MHz/div. horizontal, 10 dB/div. vertical

<table>
<thead>
<tr>
<th>Specification</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal bandwidth</td>
<td>60 MHz</td>
</tr>
<tr>
<td>Peak SNDR</td>
<td>40 dB</td>
</tr>
<tr>
<td>SFDR (at peak SNDR)</td>
<td>49 dB</td>
</tr>
<tr>
<td>Clock freq.</td>
<td>520 MHz</td>
</tr>
<tr>
<td>OSR</td>
<td>4.3</td>
</tr>
<tr>
<td>Process</td>
<td>0.25 μm digital CMOS</td>
</tr>
<tr>
<td>Area</td>
<td>1.5 mm²</td>
</tr>
<tr>
<td>Power</td>
<td></td>
</tr>
<tr>
<td>DAC</td>
<td>10 mW</td>
</tr>
<tr>
<td>Error amp.</td>
<td>100 mW</td>
</tr>
<tr>
<td>Comparators</td>
<td>30 mW</td>
</tr>
<tr>
<td>Clock buffer</td>
<td>40 mW</td>
</tr>
<tr>
<td>Shift register</td>
<td>40 mW</td>
</tr>
<tr>
<td>Total</td>
<td>220 mW (88 mA @ 2.5 V)</td>
</tr>
</tbody>
</table>

5.7 ALTERNATIVE ARCHITECTURES

The FIR DAC sigma-delta ADC fabricated is a replacement for conventional Nyquist-rate ADC’s. However, sigma-delta ADC’s may be more attractive with alternative system architectures. For radio receivers, other authors have suggested digitising the intermediate
Figure 5.12: Single-tone measurement taken using the 128 tap on-chip shift register. The 128 point FFT does not have enough frequency resolution for a meaningful two-tone test.

frequency (IF) signal with a bandpass sigma-delta ADC (Pellon 1998) (Gao & Snelgrove 1998). This avoids baseband I/Q mismatch and flicker noise problems, although sensitivity to ADC clock jitter is greater. It is more attractive for oversampling converters than Nyquist-rate converters as the clock rate need not be increased – only the passband where the NTF is low needs to be moved. Section 5.7.1 considers how to achieve this with the FIR DAC architecture.

There is also the idea of a “software radio” in which almost all signal processing is digital (and preferably in software). The FIR DAC sigma-delta ADC is interesting in this context as the loop filter is more digital in nature. Section 5.7.2 considers a variant in which the loop filter is a digital filter rather than a FIR DAC.

In light of the performance achieved by the active filter in chapter 3, it is also worth examining the possibilities with continuous-time loop filters. This is discussed in section 5.7.3

175
5.7.1 Bandpass option

Only minor changes would be required to change from a low-pass to a band-pass NTF. The natural centre frequency is \( f_s/4 = 130 \text{MHz} \), as a symmetrical response around that frequency allows half of the FIR DAC taps to be omitted. Most significantly, the troublesome first tap is not required; the loop filter is of the form

\[
L(z) = l_2z^{-2} + l_4z^{-4} + l_6z^{-6} + \cdots
\]

This would allow the sample rate to be increased somewhat, or the converter made more resistant to process variation at \( 520 \text{Ms/s} \). The only down-side is that the \( \frac{dv}{dt} \) in the error amplifier due to the continuous-time input signal being compared with discrete-time feedback would increase. This should not be a limiting factor however as the track and hold only requires around \( 30 \text{dB} \) dynamic range; at worst the track and hold may need to be moved closer to the input of the error amplifier. With a little more work on the existing imperfections (offset averaging in the quantiser and reduced error amplifier distortion by removing the first tap problem) to increase the SNDR by about \( 10 \text{dB} \), and an NTF which gave \( 10 \text{dB} \) more in-band noise reduction in return for an OSR of six rather than four, the converter could clock at \( 800 \text{Ms/s} \) and digitise a bandwidth of \( 70 \text{MHz} \) centred at \( 200 \text{MHz} \) with \( 60 \text{dB} \) SNDR. This IF digitisation would be quite attractive for software radios and wireless LAN’s.

5.7.2 Digital filter

In the ADC designed, the FIR DAC is used to shape the NTF. There is another option, namely digital filtering. That is, rather than a quantiser and a FIR DAC in the feedback loop, there would be the quantiser, some digital signal processing, and a conventional DAC (2B, fig. 5.1). As with the FIR DAC architecture, an input pre-filter before the feedback would be needed.

This is a promising architecture, as it follows the trend towards increasing use of digital signal processing (DSP) in place of analog signal processing which is expected as CMOS feature size is reduced. Using the linear predictor interpretation in chapter 4, the analog
pre-filter provides an input signal which is predictable. The quantiser, DSP and DAC then predict the next sample of the filtered input signal. Subtracting this prediction allows the quantiser input to be kept small (say, -30 dB) so that it can be quantised accurately with a low-resolution (say, four-bit) quantiser. A software radio built using this approach would be an interesting proposition.

For the converter designed, the aim was to achieve a clock rate and bandwidth beyond that possible with switched-capacitor sigma-delta ADC’s. This made the use of DSP less attractive. The latency limitations derived in chapter 4 limit pipelining. An FIR or IIR (infinite impulse response) digital filter could be built at 520MHz — witness what microprocessor arithmetic units can do — but it would be a major project in itself. About the only approach which seemed simple and practical enough to work at 520 MHz was if a simple ROM-based state machine could be used.

A ROM-based state machine seemed possible, so it was investigated further. (In practice, a static RAM would probably have been used in place of a ROM to allow programmability.) The four bits (or so) from the quantiser and \( n \) bits of state would address the ROM, from which it would calculate a DAC output of around eight bits, and \( n \) bits for the next state. The ROM would therefore require \((n + 8)2^{(n+4)}\) bits of storage. The question was whether \( n \) would be small enough for this to be practical.

**ROM size calculation**

The ROM size can be calculated using the state-space dynamic range theory in chapter 2, as the problem is that of realising a filter, \( L(z) \), with sufficient dynamic range. The discrete-time version of the theory in which a delay \( z^{-1} \) replaces an integrator is used. This has not been presented, but it is practically identical except that \( z \) replaces \( s \) and quantities such as Gramians are integrated over the unit circle \( z = e^{j\theta} \) rather than the \( j\omega \) axis (Thiele 1986). The main effort before using this theory is to calculate the input signal to the filter \( \Phi(z) \), and what weight \( \Psi(z) \) to give output noise.

Assume that the modulator uses the same NTF and four-bit quantiser as the FIR DAC realisation. For calculation purposes, assume the quantiser and DAC have unity gain, and
Using the linear NTF analysis, the input to the loop filter is

\[ U(z) = \frac{1}{1 + L(z)} [R(z) + E(z) - D(z)] \]

Assume the DAC full-scale output is ±1. The maximum input will then be around ±0.5; assume that \( R(z) \) is 0.25 RMS. The power spectral density is therefore 5 x 0.25^2 = 0.31, assuming an oversampling ratio of 5. (All spectral densities here are with respect to 2π rad, as we want a unity power white noise to give a weight of one.)

The NTF 1/(1 + L(z)) averages around -26 dB (0.05) in-band, so the peak error \( R(z)/(1 + L(z)) \) will be of the order of 0.5 x 0.05 = 0.025. The quantiser needs to cope with this and some amplified out-of-band quantisation noise, so the quantiser range will be around ±0.05. The quantiser step is therefore 0.1/16 = 0.00625. Hence the quantisation noise \( E(z) \) is 0.0031 peak, or 0.0018 RMS. This corresponds to a power spectral density of 0.0018^2 = 3.2 x 10^{-6}. We therefore choose the input weight as

\[
\begin{align*}
\left| \phi(e^{j\theta}) \right|^2 &= \frac{0.31}{[1 + L(e^{j\theta})]^2}, |\theta| \leq \frac{\pi}{5} \\
&= \frac{3.2 \times 10^{-6}}{[1 + L(e^{j\theta})]^2}, |\theta| > \frac{\pi}{5}
\end{align*}
\]

Now consider the filter rounding noise. The output is

\[ Y(z) = \frac{1}{1 + L(z)} D(z) + \frac{L(z)}{1 + L(z)} (R(z) + E(z)). \]

Assume the allowable in-band error contribution from filter rounding \( D(z)/(1 + L(z)) \) is 0.001 (-60dB). Out-of-band, \( D(z)/(1 + L(z)) \) must be small enough that it does not
overload the quantiser. Assume that 0.01 (−40 dB) is acceptable — remember the maximum quantiser input is ±0.05. If we take the output weight as

$$\left| \Psi(e^{j\theta}) \right|^2 = \frac{10^6}{|1 + L(e^{j\theta})|^2}, |\theta| \leq \frac{\pi}{5}$$

$$= \frac{10^4}{|1 + L(e^{j\theta})|^2}, |\theta| > \frac{\pi}{5}$$

then the maximum weighted output noise $\Psi(z)D(z)$ allowed is unity.

We next calculate the frequency-weighted Hankel singular values using these weights and the desired transfer function $L(z)$. For convenience, we use the same loop filter transfer function as that of the FIR DAC architecture. We only have the magnitudes of the weights not their transfer functions, but we can still calculate the Gramians according to their definitions (2.10) and (2.11). (The only change for discrete-time is to integrate around the unit circle rather than along the $j\omega$ axis.) This requires the gain to states $F(z)$ and the noise gain $G(z)$ and hence a state-space realisation, which we obtain simply from the FIR structure. We then use the property that the Hankel singular values $\sigma_i$ are the same for all realisations.

This results in

$$\sigma_i^2 = \lambda_i(KF^\Psi W^\Psi) = 2740, 230, 5.2, 0.091, \ldots.$$ 

The other eight $\sigma_i^2$ are smaller than 0.02. We recall from section 2.9.3 that for the input-normal realisation, with unit noise $e$ added to each state, the frequency-weighted output noise power from the $i$th state is $\sigma_i^2$. We also recall that the frequency-weighted input-normal realisation has a property of minimising the number of bits of storage required in a digital filter (Mullis & Roberts 1976).

We said above that the maximum weighted output noise allowed is unity. We therefore need to keep the rounding noise power contribution to each state below $1/\sigma_i^2$. (Actually it needs to be a little lower, as the noise powers from different states add.) For the last nine states, $\sigma_i$ is less than unity. This means that even if we round those states to zero, the output noise will be tolerable. The rounding noise power decreases by a factor of four for each bit of precision used to store a state. For the first three states, roughly 6, 4 and 2 bits of precision respectively will give sufficiently low rounding error. Thus the total information
storage in the state is 12 bits. The ROM must take 12 bits of state and 4 bits of input, and from these calculate 12 bits of state and around 8 bits of output. The total ROM size is therefore \((12 + 8) \times 2^{(12+4)} = 1.3 \times 10^6\).

This memory size is impractical for an embedded memory, although not hugely so. SRAM bit area in 0.18\,\mu m logic CMOS is around \(5 (\mu m)^2\), and ROM is presumably around \(1 (\mu m)^2\), so the area needed is \(6.5 (mm)^2\) in SRAM or \(1.3 (mm)^2\) in ROM. Of course, this would change greatly if margin for overload required a few more bits, or if a better transfer function allowed a few less bits. A large ROM may also be unattractively slow.

5.7.3 Continuous-time loop filter versus FIR DAC

As mentioned earlier in the chapter, the FIR DAC was used because the author did not believe that a continuous-time loop filter would be accurate enough at hundreds of megahertz. Given the 800MHz filter simulation results at the end of chapter 3, using a continuous-time loop filter looks more promising.

The FIR DAC approach has advantages and disadvantages compared to a continuous-time loop filter. The advantages are that the FIR DAC transfer function is well controlled, and that the FIR DAC output is only used once it has settled, so DAC glitches are not significant like they are when using a DAC in continuous-time. This also reduces sensitivity to clock jitter. The disadvantages are that the FIR DAC modulator requires a discrete-time amplifier which settles in around 1\,ns, and that a continuous-time input is compared with discrete-time feedback, leading to a high \(\text{dv/dt}\) for the track and hold. A continuous-time loop filter also allows slightly better shaping of the NTF – the FIR DAC, like all short FIR filters, has a broad transition band. A combination of a FIR DAC and conventional loop filter may be attractive – this approach has been tried in switched-capacitor technology by Okamoto, Maruyama & Yukawa (1993).
5.8 SIGMA-DELTA VERSUS NYQUIST-RATE CONVERTERS AT 20-200MHZ BANDWIDTH

5.8.1 Performance comparison

A signal to noise and distortion ratio of 40 dB over 60 MHz bandwidth does not sound particularly impressive. However, it was equal to the state of the art in CMOS when this work was published, and is still close. (Bipolar ADC’s achieve better performance, but offer less potential for system-on-a-chip integration.) At these bandwidths, the competitors are Nyquist-rate converters. These require an anti-alias filter. An anti-alias filter similar to that in chapter 3 with the passband edge at 60 MHz, has a stopband edge around 100 MHz. A 160Ms/s Nyquist-rate converter would therefore be required to avoid aliasing. The pre-filter required by the FIR DAC sigma-delta architecture allows the full 60 MHz bandwidth to be used – the pre-filter would roll off above 60 MHz and needs to be −30 dB or so by 90 MHz to avoid overloading the modulator.

The following are the best CMOS ADC’s the author could find for sample rates ≥100Ms/s. Note how the state of the art is without exception recent work.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Topology</th>
<th>Speed</th>
<th>SNDR</th>
<th>SFDR</th>
<th>Power</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Harrison &amp; Weste 2000)</td>
<td>FIR DAC  ΣΔ</td>
<td>“160 Ms/s”</td>
<td>40dB</td>
<td>49dB</td>
<td>220 mW</td>
<td>0.25 μm</td>
</tr>
<tr>
<td>(Choe et al. 2001)</td>
<td>folding</td>
<td>100Ms/s</td>
<td>35dB</td>
<td>45dB</td>
<td>165 mW</td>
<td>0.5 μm</td>
</tr>
<tr>
<td>(Park et al. 2001)</td>
<td>pipeline</td>
<td>100Ms/s</td>
<td>57dB</td>
<td>64dB</td>
<td>180 mW</td>
<td>0.18 μm</td>
</tr>
<tr>
<td>(Wang &amp; Razavi 2000)</td>
<td>subranging</td>
<td>150Ms/s</td>
<td>40dB</td>
<td>43dB</td>
<td>305 mW</td>
<td>0.6 μm</td>
</tr>
<tr>
<td>(Nagaraj et al. 2000)</td>
<td>flash</td>
<td>200Ms/s</td>
<td>41dB</td>
<td>45dB</td>
<td>143 mW</td>
<td>0.25 μm</td>
</tr>
<tr>
<td>(Sumanen et al. 2001)</td>
<td>4 × pipeline</td>
<td>200Ms/s</td>
<td>46dB</td>
<td>55dB</td>
<td>280 mW</td>
<td>0.5 μm</td>
</tr>
<tr>
<td>(Choi &amp; Abidi 2001)</td>
<td>flash</td>
<td>1025 Ms/s</td>
<td>36dB</td>
<td>45dB</td>
<td>500 mW</td>
<td>0.35 μm</td>
</tr>
</tbody>
</table>

These SFDR’s and SNDR’s are for similar input frequencies. The SNDR and SFDR for the ADC described is for 38 MHz / 40 MHz two-tone input, with energy above 60 MHz ignored. (A sigma-delta will have lots of out-of-band noise and possibly harmonics, but these are removed by digital filtering.) An 8MHz single-tone input (fig. 5.12), putting the
harmonics in-band, gave similar noise and distortion. Data from the other papers for
40–53 MHz input is used, except (Choe et al. 2001) which uses 12 MHz input, and
(Sumanen et al. 2001) uses 71.3 MHz input (although the converter is not significantly
better at 40 MHz).

The recent work by Sumanen et al. (2001) is the first to clearly outperform the ADC
described here. Sumanen et al. (2001) use a 200 Ms/s track and hold and four 50 Ms/s
parallel pipelines. This is quite elegant as track and holds do not require as precise a gain
as pipeline stages and hence can use lower gain, faster opamps. Bootstrapped switches and
digital correction are also used.

Unless it was an application which particularly favoured a four-bit output at 520 Ms/s
and had a readily available clock at that frequency, the sigma-delta approach would
probably not be preferable. Software radios or wireless LAN’s come to mind as an
application where the higher clock rate might be available as a local oscillator, and might
reduce spuri problems. The four-bit output might be convenient if pin count limitations
were otherwise going to require a separate high-speed serial I/O. Note that Sumanen et al.
(2001) require an additional 125 mW for their digital I/O, which is not included in the
280 mW above. The digital filtering required by sigma-delta ADC’s would probably also be
already present for radio applications.

The more relevant question is therefore whether the FIR DAC architecture has
advantages which will allow future implementations to outperform Nyquist-rate converters.

5.8.2 Limitations of sigma-delta ADC’s and Nyquist-rate ADC’s

The idea that sigma-delta modulators with continuous-time loop filters should be attractive
as ADC’s at bandwidths of tens of megahertz has been around for a number of years.
Digitising at intermediate frequency (IF) with a bandpass modulator, rather than at
baseband with a Nyquist-rate converter, is particularly mentioned (Gao & Snelgrove 1998).
Unfortunately, continuous-time modulators have never really lived up to expectations, with
degradations attributed to loop filter dynamic range limitations and poor signal scaling,
clock jitter (Cherry, Snelgrove & Schvan 1997) and metastability. Should such modulators
be written off, or is just a little more attention to detail needed?

**Limitations of pipelined converters**

It seems reasonable to look at sigma-delta modulators only if Nyquist-rate converters will hit fundamental performance limits which can be circumvented by sigma-delta modulators. Pipelined converters seem the most attractive architecture for achieving bandwidths of 50-100 MHz at 60-80 dB SNDR, so it is worth checking what limits they face. Folding and flash converters seem to have more fundamental limitations relating to transistor mismatch and complexity which make them less likely to jump from the 8-bit level to the 12-bit level, although they will presumably still be competitive at higher sample rates.

The fundamental limitations of pipelined converters appear to be that the sampling rate is limited by opamp settling time, and that the SNDR for a given power consumption is limited by thermal noise. ADC's using bipolar transistors offer some possibility to get high dynamic range without feedback, but with deep submicron CMOS and its inferior transistor gain, feedback around amplifier transistors seems inevitable.

Existing switched-capacitor opamps, even in deep submicron CMOS, achieve gain-bandwidths around 1 GHz (Mehr & Singer 2000) (Park et al. 2001) and settling times around 4 ns. Speed might be expected to scale with transistor $f_T$, but it tends to fall short because it is harder to operate at maximum $f_T$ at lower $V_{DD}$, and the stage gain of minimum-length transistors is falling.

The author has tried minimising the settling time of variants of the opamps in chapter 3. The results were similar to those achieved with existing techniques such as Miller-compensated opamps. (Gain-boosted cascodes (Bult & Geelen 1990) tend to be very hard to implement at low $V_{DD}$.) If power consumption is not an issue, settling times of 1-2 ns are possible. Once power consumption is restricted and process variations are included, fast settling with the capacitive loads needed for low thermal noise is difficult. Fast settling is more difficult than wide bandwidth because optimum phase margin is higher ($60 - 70^\circ$), and because doublets must be pushed to high frequencies (Bult & Geelen 1990). Stages in a pipelined converter also run at a gain of two or more, not unity-gain.
Thermal noise advantages of sigma-delta modulators

The main advantage the author sees for sigma-delta ADC’s with continuous-time loop filters is that they should be able to achieve lower thermal noise. Thermal noise formulae are in (Goes, Vital & Franca 1998) for pipelined ADC’s, (Dias, Palmisano, O’Leary & Maloberti 1992) for switched-capacitor sigma-delta ADC’s, and (Dias, Palmisano & Maloberti 1992) for continuous-time (and mixed continuous-time / discrete-time) sigma-delta ADC’s.

Kelly, Yang, Mehr, Sayuk & Singer (2001) state that the first stage of a practical pipelined converter needs 4 pF capacitors for 70 dB SNR. This is with a multi-bit first stage so that the noise of later stages is less significant, and 2 Vp-p differential signal swing (using 3 V $V_{DD}$). The opamp for the stage with 4 pF capacitors consumed 100 mW to get fast enough settling for 75 Ms/s. With lower power supply voltages and the single-bit-per-stage pipelines needed for maximum sample rate it only gets worse.

The thermal noise of a sigma-delta modulator with continuous-time loop filter is largely determined by the first integrator (Dias, Palmisano & Maloberti 1992). For 50 MHz signal bandwidth and 4 pF integrator capacitors, the loop filter bandwidth is likely to be around 100 MHz, resulting in input resistors around 400 Ω. If the input signal is 1 Vp-p differential, the input current is 2.5 mAp-p, requiring a DAC tail current around 5 mA. Such a DAC is likely to have transistor $g_m$ of around 5 mS per side. The total thermal-noise-producing conductance at the input is 2.5 mS from the input resistor, 5 mS from the DAC (the theoretical value is $\frac{2}{3}g_m$, with extra in practice from hot carriers), and perhaps another 5 mS from the opamp and other resistors in the filter. This 12.5 mS contributes 100 nA RMS noise over 50 MHz bandwidth, or 140 nA RMS counting both sides. With respect to the 2.5 mAp-p input signal, this is an SNR of 76 dB. The opamp output current depends on how well the DAC output matches the input signal and on second-stage loading, but 5 mA output-stage tail current (assuming a similar architecture to that in chapter 3) seems reasonable. The total opamp power is likely to be around 10 mA or 20 mW. This clearly looks more promising than the pipelined converter.

FIR DAC architectures are also potentially quieter than pipelined architectures, as the
error signal can be amplified before sampling. The signal is also only sampled once and does not accumulate multiple $kT/C$ noises like that in a pipelined converter.

**Limitations of sigma-delta modulators**

Having observed that pipelined converters face fundamental thermal noise and settling time limits, and that sigma-delta modulators offer advantages with respect to thermal noise, we should check that sigma-delta modulators do not have other worse fundamental limitations of their own. Clock jitter, metastability and loop filter dynamic range have been raised as significant issues (Cherry & Snelgrove 1999a) (Tao, Tóth & Khoury 1999) (Gao & Snelgrove 1998). If a multi-bit DAC is to be used, DAC resolution is another significant consideration.

Loop filter dynamic range is calculated in the example above. We saw in section 2.4 that the noise of an opamp-RC filter is mostly that of its resistors (and DAC in a sigma-delta ADC). Bandpass modulators could pose greater difficulties, because of the dynamic range reduction experienced in accordance with the dynamic range limitation of section 2.6. Note that anti-alias filters are potentially a bigger dynamic range problem than sigma-delta modulator loop filters, as the dynamic range required of loop filters is reduced by being in the feedback loop.

Metastability can be reduced to acceptable levels by using one or more latches after the comparator. The fundamental limitations with respect to excess loop delay in chapter 4 are relevant here. However, the $f_T$ of deep submicron CMOS is very high, so 1–2 Gs/s seems achievable with negligible metastability.

Clock jitter is a particular problem if continuous-time loop filters are used. (A switched-capacitor or FIR DAC sigma-delta ADC has no disadvantage compared to a Nyquist-rate ADC – they all sample the input signal. The FIR DAC architecture samples the error rather than the input, but as the DAC pulse being subtracted has settled there is no difference.) With a continuous-time loop filter, the clock is used in two places, the quantiser and the DAC. The quantiser clock is non-critical, as the quantiser resolution is low. The DAC clock is critical.
Figure 5.14: Typical time-domain output of the FIR DAC sigma-delta ADC. (Data collected at 520 MHz, then clocked out at 1 kHz from the on-chip shift register.) The high-frequency noise is due to the high out-of-band NTF.

Clock jitter is analysed in Cherry & Snelgrove (1999a) and Tao et al. (1999). Cherry & Snelgrove (1999a) conclude that it need not constrain performance if a conventional on-chip LC oscillator is used. This is for one-bit quantisers, so we should check that multi-bit quantisers and high out-of-band NTF’s do not cause problems.

Clock jitter can be analysed using the NTF analysis. The gain from the DAC to the output is equal to the signal transfer function. In-band (where the jitter noise matters), this is close to unity. The magnitude of the jitter noise input basically depends on the average change in DAC output each clock cycle (Cherry & Snelgrove 1999a). For a one-bit quantiser, this is large as the output is either +1 or −1, and alternates rapidly between the two. For a low-pass modulator with multi-bit quantiser and low out-of-band gain, the DAC change in output is usually only a small fraction of full scale, so clock jitter is less of a problem. Using a multi-bit quantiser with high out-of-band gain, the change in output is a significant fraction of full scale – see fig. 5.14. As this is somewhat better than the one-bit case, which Cherry & Snelgrove (1999a) do not consider a problem, we can conclude that
clock jitter will not fundamentally limit performance. Certainly care needs to be taken to avoid digital data disturbing the clock however.

The other major challenge with such a sigma-delta modulator is to build a 1–2 Gs/s DAC with sufficient SNDR. Decoding-induced delay variations are not a problem as the natural approach is to put the thermometer outputs from the quantiser straight into the DAC. (The FIR DAC converter also used this approach.) Signal-dependent glitches are more of a problem, although return-to-zero DAC pulses could be used. Van den Bosch, Borremans, Steyaert & Sansen (2001) describe a DAC with 70 dB SFDR at 1 Gs/s for 100MHz input without return-to-zero, so thermal-noise-limited performance is not too far away. Only having 16 or so current cells and having the virtual ground of an opamp input to work into should help.

5.9 CONCLUSION

The FIR DAC architecture is competitive with Nyquist-rate ADC's at wide bandwidths, but needs further development to be of real interest to industry. The future of it and other wide bandwidth sigma-delta modulators depends a lot on the extent to which pipelined converters improve in the future. The FIR DAC architecture and other sigma-delta ADC's seem to have an advantage with respect to thermal noise.

The combination of a multi-bit quantiser and high out-of-band NTF to achieve high resolution at low OSR seems useful. It is difficult to decide whether a continuous-time or FIR DAC loop filter is superior for clock rates beyond the capacity of switched-capacitor circuits. The FIR DAC architecture is a little inelegant in that it subtracts a discrete-time signal from a continuous-time signal. With low-pass modulators loop latency problems also limit speed; fortunately with $f_s/4$ bandpass modulators this is far less of an issue as an extra cycle is available. The FIR DAC architecture avoids problems with clock jitter and DAC nonlinear glitch energy suffered by modulators with continuous-time loop filters.

The variant in which the loop filter is fully digital has conceptual elegance. The use of a DAC to subtract linearly predicted interferers before the signal is quantised seems a good
prospect for a practical software radio. At present, the information storage is prohibitive for a ROM-based state machine, but this may change with future digital CMOS scaling.
6 CONCLUSION AND SUGGESTIONS FOR FUTURE WORK

6.1 CONCLUSION

Let us return to two questions posed in the introduction. Why do high-Q filters composed of integrators tend to have poor dynamic range? High-Q filters have poor dynamic range because the transfer function \( H(s) \) maps closely spaced points \( s_1 \) and \( s_2 \) in the \( s \)-plane to far away points \( H(s_1) \) and \( H(s_2) \) in the \( H(s) \)-plane. This is an ill-conditioned operation, which requires taking the difference of nearly equal quantities. The mathematical description of this is the dynamic range limitation

\[
\frac{H(s_1) - H(s_2)}{s_2 - s_1} = G(s_1)F(s_2).
\]

We saw that high dynamic range required making the gain to states \( F(s) \) and the noise gain \( G(s) \) small.

The dynamic range limitation is a particularly useful result because when considered over all frequencies \( \omega_1 \) and \( \omega_2 \), the limitation is achievable, namely by the frequency-weighted input-normal realisation or frequency-weighted balanced realisation. Unfortunately, this realisation is difficult to implement because of the dense coefficient matrices \((A, B, C, D)\). However, we saw in chapter 3 that practical realisations from simulations of LC ladders had very similar performance. This means that the dynamic range limitation is still useful as an accurate guide to the choice of transfer function. Low-ripple Elliptic transfer functions work well for filters.

A second question asked how sigma-delta modulators traded bandwidth for dynamic range. The existing explanation that the feedback loop corrects for errors is reasonable. Another explanation is that the modulator predicts the next sample, and hence only needs to quantise a small error in the prediction rather than the complete input. This reduces the number of bits required in the quantiser. This linear predictor interpretation seems
particularly applicable to the sigma-delta ADC in chapter 5, where the input signal is pre-filtered to make it predictable.

The question then arises as to how much feedback is possible, or how accurately the input signal can be predicted. We attempted to answer this in section 2.10.1 for amplifiers, using the dynamic range limitation, but the results were numerical rather than analytic. Optimising the amount of opamp loop gain was also an important part of the filter design in chapter 3. This was the major innovation which allowed the filter to operate at 350MHz. Again, the approach was empirical. For the sigma-delta modulators in chapter 4, we have a stronger result however, namely the Bode integral. This states that the noise transfer function in decibels integrates to zero over frequency. As the noise transfer function is constrained out of band by stability, we can accurately calculate how much quantisation noise reduction is possible.

The Bode integral states that the oversampling ratio of a sigma-delta modulator can only be reduced by increasing the out-of-band gain or sharpening the loop filter rolloff. The fact that increased out-of-band gain is possible with a multi-bit quantiser was the key to achieving a wide bandwidth from the ADC in chapter 5. This benefit of multi-bit quantisers has been known previously to some degree, but it becomes a lot clearer with the Bode integral that high out-of-band gain is vital for maximum bandwidth.

The relevance of the Bode integral has been identified by previous authors, but a second integral concerning excess loop delay is new. It makes the prediction that a sample of excess loop delay is necessarily detrimental (to the extent of halving the bandwidth) for low-pass modulators. This harks back to interpretation of a sigma-modulator predicting its future input. If the recent past is unavailable, the predictions are less accurate.

To finish, a few words on the effect of scaling on analog CMOS. The author’s belief is that the greater use of feedback possible with faster transistors will more than counter the falling swing as a percentage of supply voltage down to at least $V_{DD} = 1$ V. Low-threshold transistors may be required, but microprocessors use these too. Today’s 10 MHz gm-C filter will be successively replaced by the class-A opamp-RC filter and the class-B opamp-RC filter. With more loop gain, output stages can be driven closer to cutoff and saturation for
the same distortion. Cascodes are no longer needed. Loop gain is not going to fall because the transistor stage gain falls; the feedforward techniques in chapter 3 allow three and four stage opamps to operate at wide bandwidth. Pipelined ADC’s will have the bandwidth of today’s flash ADC’s, and sigma-delta ADC’s will have the bandwidth of today’s pipelined ADC’s. The techniques in this work are adaptable to at very least the 0.13μm process generation. It should be emphasised that this is the author’s opinion; this work is about the state-space and sigma-delta modulator dynamic range theories, not scaling.

6.2 SUGGESTIONS FOR FUTURE WORK

6.2.1 Circuit design

Feedback amplifiers

The improvement of opamp bandwidth from 1 GHz to 10 GHz achieved in chapter 3 suggests that opamps should be used more at high frequencies. Active filters and intermediate-frequency amplifiers are obvious applications, but what could be more interesting is the use of feedback amplifiers at RF in the commercially important 900 MHz – 2.4 GHz bands. Feedback amplifiers potentially allow higher IP3 and tightly-controlled gain. Feedback amplifiers tend to become practical at frequencies around \( f_T/100 \), although this can be pushed to \( f_T/30 \) or so (as in the filter at the end of chapter 3). Transistor \( f_T \) is entering this range with the 0.18-0.13μm generations.

Feedback amplifiers at RF face peculiar design challenges. The input impedance should be resistive (eg. 50 Ω) rather than an open circuit. Inductive sources and loads present stability challenges. It may be possible to deal with these using a combination of shunt and series feedback, and Zobel networks to resistively-terminate the feedback amplifier out-of-band.

Sigma-delta modulators

As mentioned in chapter 5, a sigma-delta modulator with continuous-time loop filter using the opamp-RC technology in chapter 3 appears promising. The opamp-RC technology gives
A controlled transfer function at up to 1 GHz, which should allow high out-of-band NTF (with a multi-bit quantiser) to enhance performance. An \( f_s/4 \) bandpass version of the FIR DAC architecture was also highlighted as worth investigating, as it avoids latency problems.

### 6.2.2 Circuit theory

The frequency-domain state-space realisation theory in Chapter 2 seems fertile. There is work to be done in converting the time-domain state-space realisation results of Kalman and coworkers into the frequency-domain. The theory also may be helpful for deriving the long sought-after (Anderson & Moore 1989) (Zhou 1996) frequency-domain error bound for frequency-weighted balanced truncation.

**Linear Fractional Transformations**

A generalisation of the state-space equations is the linear fractional transformation

\[
\begin{align*}
\mathbf{u} &= \mathbf{Ax} + \mathbf{Br} \\
\mathbf{y} &= \mathbf{Cx} + \mathbf{Dr} \\
\mathbf{x} &= \mathbf{Su} + \mathbf{e}.
\end{align*}
\]

Here \( r \) is the system input, and \( y \) is the system output. In a state-space filter, \( S \) would be a diagonal matrix of integrators, and \( A, B, C \) and \( D \) would be frequency-independent. These equations result in a gain from input \( r \) to input to the integrators \( \mathbf{u} \) of \( \mathbf{F} \), and a gain from added noise \( \mathbf{e} \) to output \( \mathbf{y} \) of \( \mathbf{G} \), where

\[
\begin{align*}
\mathbf{F} &= (\mathbf{I} - \mathbf{AS})^{-1} \mathbf{B} \\
\mathbf{G} &= \mathbf{C}(\mathbf{I} - \mathbf{SA})^{-1}.
\end{align*}
\]

The overall transfer function from \( r \) to \( y \) is

\[
H = \mathbf{CS}(\mathbf{I} - \mathbf{AS})^{-1} \mathbf{B} + \mathbf{D} = \mathbf{C}(\mathbf{I} - \mathbf{SA})^{-1} \mathbf{SB} + \mathbf{D}
\]

An equivalent of the dynamic range limitation can be derived. If \( A, B, C \) and \( D \) are frequency independent, and \( S = S(j\omega) \),

\[
H(j\omega_1) - H(j\omega_2) = \mathbf{G}(j\omega_1)[S(j\omega_1) - S(j\omega_2)]\mathbf{F}(j\omega_2).
\]
If this could somehow be given a singular value decomposition in analogy to the work in chapter 2, we would have a synthesis which allows a transfer function to be realised from an arbitrary resonator $S(j\omega)$. There may also be possibilities for circuit-level synthesis; linear fractional transformations are the basis of the Youla (1961) cascade synthesis variant of Darlington synthesis. Linear fractional transformations are also widely used in control theory (Zhou 1996).


Bult, K. (2000), Analog design in deep sub-micron CMOS, in ‘European Solid-State Circuits Conf.’, number 305.


195


IEEE Std. 802.11a (1999), ‘Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: High-speed physical layer in the 5GHz band’.


201


