Classification and Design Space Exploration of Low-Power Three-Stage Operational Transconductance Amplifier Architectures for Wide Load Ranges

Joseph Riad 1,*, Johan J. Estrada-López 1,2,* and Edgar Sánchez-Sinencio 1

1 Electrical and Computer Engineering Department, Texas A&M University, College Station, TX 77843, USA; s-sanchez@tamu.edu
2 Faculty of Mathematics, Autonomous University of Yucatán, Mérida, Yucatán 97000, Mexico
* Correspondence: joseph.riad@tamu.edu (J.R.); johan.estrada@tamu.edu (J.J.E.-L.)

Received: 29 September 2019; Accepted: 29 October 2019; Published: 1 November 2019

Abstract: Since operational transconductance amplifiers (OTAs) form the basic building blocks of many analog systems, the compensation of three-stage OTAs has attracted a lot of attention in the literature. Many different solutions to the stability problem of such OTAs have been proposed over the past 20 years, with each solution exhibiting different properties or targeting a different application. This work surveys a broad selection of previously reported architectures and proposes a novel classification scheme that exposes features common to seemingly different compensation architectures and serves as a guideline for which type of OTA is suitable for a given application. In addition, a novel figure of merit (FoM) is proposed to guide the designer in deciding which OTA architecture suits the tradeoffs specific to the application at hand. Theoretical discussions are further reinforced by transistor-level simulation results.

Keywords: three-stage OTA; Miller compensation; nested Miller compensation; reverse nested Miller compensation; wide load range

1. Introduction

With the continued aggressive scaling of CMOS technologies, the speed of digital circuits has been increasing and their power consumption decreasing. These benefits, however, come at the cost of decreased intrinsic device gain, thereby hurting the performance of analog circuits [1]. On the other hand, high-precision applications require operational transconductance amplifiers (OTAs) with high DC gain. The traditional solution for this problem has been to stack transistors vertically in a cascode configuration so as to achieve a high DC gain with a nearly first-order response.

As scaling continues, however, supply voltages also need to be scaled down to ensure device reliability. This results in reduced signal headroom and renders conventional cascode techniques unreliable [2,3]. For this reason, modern OTAs tend to involve a cascade of multiple stages (three or more) to achieve the desired gain.

Even though some authors have demonstrated the successful design of four-stage OTAs [4–6] and some have gone beyond that to describe \( n \)-stage OTAs [7–9], three-stage OTAs remain a reasonable tradeoff between complexity and power efficiency and have therefore garnered a lot of research interest for diverse applications over the past 20 years or more.

Applications for three-stage OTAs include headphone amplifiers, liquid crystal display (LCD) drivers, low-dropout (LDO) linear regulators and capacitive MEMS sensors [3,10–13]. Some applications (for example, MEMS and active matrix LCD) require the amplifier to drive very large...
capacitive loads [14] and others (e.g., headphone drivers and MEMS sensors) need the amplifier to be able to drive a wide range of load capacitors over several orders of magnitude [3,15].

The main challenge in three-stage OTA design is the compensation of the resulting three-pole system. The compensation architectures that have been devised to address this issue tend to be complicated and defy a tractable intuitive analysis and several works have been dedicated to deriving intuitive expressions for three-stage OTA transfer functions once their compensation structure is given [16–20]. These works allow engineers to quickly derive expressions for pole and zero frequencies to be used in hand analysis and design but do not comment on the relative merits of different compensation architectures.

Moreover, this challenging and extended design space has led to a proliferation of solutions and approaches to the compensation problem, with Miller compensation being chief among them. Classical architectures relied on nested Miller compensation [2,7,21–27] and it remains an attractive solution today [12,28]. Architectures that rely on reverse nested Miller provide an improved power efficiency [3,13,29–33] but it has been recognized that nesting Miller capacitors leads to bulky and slow implementations and many architectures were devised to use a single Miller capacitor along with some ancillary compensation structures [4,11,14,34–48]. Some authors have also demonstrated compensation techniques that do not rely on a Miller capacitor at all [8,49–52].

With the plethora of existing architectures, the task of choosing a compensation technique for a given application becomes daunting and a review of available solutions is much needed. To this end, many reviews and tutorials have been published to compare existing compensation architectures [53–55], analyze their distortion and noise performance [56–58] and optimize their settling time performance [59–63].

Most of these tutorials and reviews focus on a specific application or a specific type of compensation architecture and very few of them provide the means to compare different compensation architectures prior to having designed them at the transistor-level. In this review article, a large selection of three-stage OTA compensation architectures are reviewed and compared. Furthermore, a novel figure of merit is proposed, allowing a priori comparison of the power efficiency of different compensation architectures. This figure of merit exposes the tradeoffs involved in each compensation technique and can be used as a guide to architecture selection once an analytical expression for the OTA’s transfer function is obtained.

In addition to fine-grained architecture comparison, a taxonomic classification of the extant compensation architectures is proposed. This taxonomy divides the different compensation architectures into three broad categories and extracts the common features of the architectures in each one of them. It can therefore be used to steer the design focus towards compensation architectures that are more suitable to the application at hand and even predict qualitative properties of new OTA architectures according to where they fall in the taxonomy. To the authors’ best knowledge, this is the first time such a classification scheme has been proposed.

The rest of this paper is organized as follows: Section 2 starts by examining control-theoretic issues common to all three-stage OTAs where it is noted that the current approach of designing the amplifier for a target phase margin [3,33,43–46,48] without regard to other stability metrics can lead to a design that performs sub-optimally [62,64] and is wasteful of power. With the common challenges noted and the design procedure of optimizing the response for settling time instead is outlined, the proposed figure of merit for architecture comparison is explained in detail. Section 3 describes the proposed OTA classification scheme and describes the common features and the suitable applications for each category. Section 4 discusses circuit-level considerations that arise during the implementation of three-stage OTAs. In Section 5, transistor-level simulations of selected architectures at a fixed power budget help confirm the discussion and make it more concrete. Conclusions are given in Section 6.
2. A Control Perspective on Multistage Amplifier Design

Compared to a two-stage OTA, ensuring the stability of a three-stage OTA is more complicated since the added high-impedance node introduces an additional low-frequency pole to the transfer function.

As the transfer function order increases, several challenging issues arise in the design. This section highlights these issues and outlines an alternate design procedure to overcome them. Once the general design procedure is outlined, a method for comparing architectures and selecting a suitable one for a given application is proposed, so that the design procedure can be applied to a specific architecture.

2.1. Challenges in Multistage Amplifier Compensation

The first main design challenge is that the phase margin, used extensively as a stability metric in the design of two-stage OTAs, is no longer an adequate indicator of stability by itself. Furthermore, it is quite difficult to derive accurate phase and gain margin formulas to be used for hand analysis and design. These issues are explored in detail in the following subsections.

2.1.1. Inadequacy of the Phase Margin as a Stability Criterion

Consider a simplified three-stage OTA with a dominant pole, a pair of non-dominant poles and no zeros. The open-loop gain of such an OTA may be expressed as

\[
A_v(s) = \frac{A_0}{\left(1 + \frac{s}{\omega_{pd}}\right)\left(1 + \frac{s}{\omega_0 Q} + \frac{s^2}{\omega_0^2}\right)}
\]  

where \(A_0\) is the DC gain, \(\omega_{pd}\) is the dominant pole frequency and \(\omega_0\) and \(Q\) represent the natural frequency and quality factor of the non-dominant pole pair, respectively.

As an example implementation, consider the small-signal block diagram shown in Figure 1 which represents a nested \(g_m\)-\(C\) compensation architecture [7]. \(g_{oi}\) and \(C_i\) represent the output conductance and parasitic capacitance of stage \(i\), respectively, while \(C_L\) represents the load capacitance. With \(g_{m_{f1}} = g_{m_1}\) and \(g_{m_{f2}} = g_{m_2}\), the open-loop transfer function of this OTA has the same form as Equation (1). Using the notation of Figure 1, the parameters of this OTA’s transfer function are shown in Table 1.

**Figure 1.** Small-signal model of an example three-stage OTA.
Table 1. Transfer function parameters of the OTA shown in Figure 1.

<table>
<thead>
<tr>
<th>$A_0$</th>
<th>$\omega_{pd}$</th>
<th>$\omega_0$</th>
<th>$Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_{m1}, g_{m2}, g_{m3}$</td>
<td>$g_{m2}, g_{m3}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_m C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_m C_L}}$</td>
</tr>
</tbody>
</table>

Under unity-gain feedback, the closed-loop transfer function of this OTA becomes

$$A_{CL}(s) = \frac{A_0}{A_0 + \left(1 + \frac{s}{\omega_{pd}}\right) \left(1 + \frac{s}{\omega_0 Q} + \frac{s^2}{\omega_0^2}\right)}$$  \hspace{1cm} (2)

Applying the Routh–Hurwitz stability criterion [65] to this function yields the following stability condition

$$\left(\frac{1}{\omega_{pd}} + \frac{1}{\omega_0 Q}\right) \left(\frac{1}{\omega_{pd}} + \frac{1}{\omega_0^2}\right) > \left(A_0 + 1\right) \cdot \frac{1}{\omega_{pd} \omega_0^2}$$  \hspace{1cm} (3)

Under the assumptions $A_0 \gg 1$, $\omega_{pd} \ll \omega_0 Q$ and $\omega_{pd} \omega_0 Q \ll \omega_0^2$, this can be simplified to

$$\frac{\omega_0}{Q} > 1$$  \hspace{1cm} (4)

with $\omega_0$, the normalized non-dominant pole frequency, defined as $\omega_0 = \frac{\omega_0}{GBW} = \frac{\omega_0}{A_0 \omega_{pd}}$ where GBW stands for the gain-bandwidth product. With reference to the example of Figure 1 and Table 1, one gets

$$\frac{\omega_0}{Q} = \frac{1}{GBW} \cdot \sqrt{\frac{g_{m2} g_{m3}}{C_m C_L}}$$  \hspace{1cm} (5)

so, for a specified GBW, the parameter $\omega_0$ correlates with power consumption through the transconductances $g_{m2}$ and $g_{m3}$. This is true for other three-stage OTAs as well since $\omega_0$ relates to how far the non-dominant poles are pushed beyond the GBW.

Denoting the unity-gain frequency (UGF) by $\omega_\mu$, the phase margin may be expressed with reference to Equation (1) as

$$\phi_m = 180^\circ - \arctan\left(\frac{\omega_\mu}{\omega_{pd}}\right) - \arctan\left[\frac{1}{Q \left(\frac{\omega_0}{\omega_{pd}} - \frac{\omega_\mu}{\omega_0}\right)}\right]$$  \hspace{1cm} (6)

Moreover, approximating the UGF using GBW ($A_0 \omega_{pd}$) (this assumes, as expressed in Equation (1), a single dominant pole and therefore a 20 dB/decade magnitude roll-off from $\omega_{pd}$ to the UGF), the phase margin may be approximated as

$$\phi_m \approx 90^\circ - \arctan\left[\frac{1}{Q \left(\frac{\omega_0}{\omega_{pd}} - \frac{\omega_\mu}{\omega_0}\right)}\right]$$  \hspace{1cm} (7)

Contours for the approximate phase margin function in Equation (7) are plotted in the $\omega_0 - Q$ space in Figure 2. The figure shows clearly that, even when the phase margin is as high as $80^\circ$, some designs can violate the Routh–Hurwitz criterion and end up in the unstable region because they have negative gain margin, as demonstrated below. This means that there will be more than one UGF leading to the formula in Equation (7) no longer being valid.
Since there is an infinite number of ways to achieve a given phase margin, the inset plot compares the unity-gain feedback responses for two different designs that both have a phase margin of 60° usually deemed enough for most designs. The red plot corresponds roughly to the approach of requiring the closed loop poles to correspond to those of a third-order Butterworth filter (see, for example, [22,25]) while the blue plot corresponds to setting $\omega_0 = 2$.

The plots demonstrate the superiority of the Butterworth pole spacing approach while also highlighting the inadequacy of relying on the phase margin as the sole measure of stability with no consideration given to other metrics such as the gain margin or the Routh–Hurwitz criterion. The situation, of course, gets more complicated when the OTA has zeros and/or additional parasitic poles close to its UGF as in more complicated compensation architectures.

![Figure 2. Phase margin contours of an all-pole three-stage OTA. The inset plot shows the unity-gain feedback step responses for two different OTAs both having a phase margin of 60°.](image)

2.1.2. Difficulty of Estimating the Gain and Phase Margins

Another challenge in the design of three-stage OTAs lies in the difficulty of estimating the gain and phase margins accurately.

In the simplified case of three poles and no zeros, estimating the gain margin is relatively simple as shown below but in the general case, when the OTA has zeros and/or additional parasitic poles, estimating the gain margin accurately through hand analysis is quite difficult if not outright impossible.

In addition, estimating the phase margin accurately, even in the simple case of the OTA given by Equation (1), is quite tricky because finding the gain crossover frequency requires solving the sixth-order equation

$$A_0^2 = \left( \frac{\omega_\mu}{\omega_{pd}} \right)^2 \cdot \left\{ 1 - \left( \frac{\omega_\mu}{\omega_0} \right)^2 \right\}^2 + \left( \frac{\omega_\mu}{\omega_0 Q} \right)^2$$

(8)

This equation can be shown to reduce to the commonly-used estimate $\omega_\mu \simeq A_0 \omega_{pd}$ when $\omega_0 \gg \omega_\mu$ and $Q\omega_0 \gg 2$. When these conditions are not fulfilled, however, approximating the gain crossover frequency using the gain-bandwidth product is not accurate. This can be seen in Figure 3 where the phase margin of the same OTA is plotted as a function of $\omega_0$ for several values of $Q$. Along with the results of numerical simulations using Equation (8), the dashed plots show the phase margin estimated from the approximate formula (see Appendix A):

$$\phi_m \simeq \arctan \left[ 1 - \left( \frac{\omega_\mu}{\omega_0} \right)^2 \right]$$

(9)
along with $\omega_0 \simeq A_0 \omega_{pd}$. As the figure shows, the estimate is quite good for $Q = \frac{1}{\sqrt{2}}$ (which is the value used in the Butterworth pole spacing approach) but can be quite inaccurate for other values of $Q$ and its accuracy gets worse as $\omega_0$ is decreased in order to save power. In the worst case, the required $\omega_0$ (and therefore power) to achieve a phase margin of 45° is overestimated by 17% when using the analytical formula, leading to an over-designed solution.

![Figure 3](image_url)

**Figure 3.** The phase margin estimate loses accuracy in low-power designs: as $\omega_0$ is decreased, and depending on $Q$, relying on the simple estimate can either lead to overestimating or underestimating the required $\omega_0$ for a given phase margin by up to 17%.

### 2.1.3. Alternative Design Approach

The conventional approach used in recent designs that target wide load range applications [3,33,43–46,48] is to design the amplifier to meet a minimum phase margin specification at the maximum desired load capacitance. However, as argued above and recognized in [62], targeting a specific phase margin is not necessarily the optimal strategy for achieving a fast settling response. Furthermore, as shown below, the same phase margin may be achieved with quite different power budgets. Therefore, a good design procedure should focus on settling response parameters from the start.

The alternative approach suggested here is to focus on the time-domain settling behavior and optimize the settling time as has been suggested in [59] and the overshoot as a function of amplifier parameters.

As an example to further reinforce the point, Figure 4 shows the phase margin and the normalized settling time ($T_s \cdot GB$) and/or overshoot percentage for the unity-gain feedback step response of the OTA given by Equation (1). These quantities are plotted as functions of $Q$ for different values of $\omega_0$. As the figures show, maximizing the phase margin does not correspond to minimizing the time-domain settling parameters. Note from the figure that, even though phase margin is not a sufficient indicator of settling performance, values of $\omega_0$ where the maximum achievable phase margin is low have large values for the minimum possible settling time and overshoot, so that phase margin puts a lower limit on the best achievable settling performance [62].

To compare the results of different design approaches, the performance of four different solutions is compared. The parameters of the different designs are shown in Table 2 along with the rationale used in selecting them.
Table 2. Parameters selected for four different design cases for comparison.

<table>
<thead>
<tr>
<th>Case</th>
<th>$\omega_0$</th>
<th>$Q$</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$2\sqrt{2}$</td>
<td>0.3</td>
<td>Same $\omega_0$ as the Butterworth approach but 45° phase margin.</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.7</td>
<td>Low-power design with optimal settling time.</td>
</tr>
<tr>
<td>3</td>
<td>2.7</td>
<td>0.9</td>
<td>Less than 10% overshoot with minimum settling time.</td>
</tr>
<tr>
<td>4</td>
<td>$2\sqrt{2}$</td>
<td>$1/\sqrt{2}$</td>
<td>Butterworth pole spacing approach.</td>
</tr>
</tbody>
</table>

With reference to Table 1, the power requirements for an NGCC OTA to achieve certain $\omega_0$ and $Q$ values can be estimated. In particular, the transconductance of the last stage can be estimated as

$$g_{m3} = \frac{\omega_0}{Q} \cdot C_L \cdot GBW$$

(10)

which can then be used to compare the designs.

The four design cases in Table 2 were numerically simulated, assuming a GBW of $2\pi \times 1 \text{ Mrad}_s$, and their performance is summarized in Table 3 along with the calculated value of $g_{m3}$ required to drive a 100 pF load with the above-mentioned GBW. The table illustrates that designing the OTA for a specified phase margin can lead to a significant power dissipation (Case 1) that is more than $3 \times$ that of the design that obtains a similar performance with a focus on minimizing settling time at a given power consumption (Case 2). Furthermore, at nearly the same power expenditure, the settling time can be improved by 47.2% when the main goal is to minimize overshoot and settling time (Case 3). Finally, it is shown that this approach achieves a similar performance to the Butterworth pole spacing approach (Case 4) at 25% less power consumption.

Table 3. Performance of the four design cases given in Table 2.

<table>
<thead>
<tr>
<th>Case</th>
<th>Phase Margin (deg.)</th>
<th>Settling Time (µs)</th>
<th>Overshoot (%)</th>
<th>$g_{m3}$ (mS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45.54</td>
<td>1.41</td>
<td>22.62</td>
<td>5.92</td>
</tr>
<tr>
<td>2</td>
<td>47.87</td>
<td>1.25</td>
<td>22.96</td>
<td>1.80</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>0.66</td>
<td>8.22</td>
<td>1.88</td>
</tr>
<tr>
<td>4</td>
<td>60.49</td>
<td>0.53</td>
<td>8.15</td>
<td>2.51</td>
</tr>
</tbody>
</table>

Thus, from a performance perspective, once a compensation architecture is selected, macro-model simulations should be used to generate curves similar to Figure 4 for the selected architecture and select the OTA parameters based on the specified time-domain performance parameters at maximum load capacitance. The issue of selecting a compensation architecture is tackled next.

2.2. Architecture Selection

The addition of an extra gain stage and other compensation structures means that the design space of the OTA is quite larger than that for a two-stage OTA, and there are many more possible architectures and compensation strategies for a multi-stage OTA [2,8,31,36,37,49]. This abundance of architectures can be confusing for designers searching for an OTA architecture suitable for a given application.
For comparing different architectures, several different figures-of-merit (FoM) have been proposed. The most famous set of FoMs is (see, for example [27,38,48]):

\[
\begin{align*}
F_{OM}^{S} &= \frac{GBW \cdot C_L}{\text{Power}} \quad \text{IFOM}_{S} = \frac{GBW \cdot C_L}{I_{dd}} \\
F_{OM}^{L} &= \frac{SR \cdot C_L}{\text{Power}} \quad \text{IFOM}_{L} = \frac{SR \cdot C_L}{I_{dd}}
\end{align*}
\]

(11)

where GBW, SR, \(C_L\) and \(I_{dd}\) denote the gain-bandwidth product, the slew rate, the load capacitance and the supply current, respectively. These FoMs are widely used because they capture both the small- and large-signal settling behavior of the OTA while also favoring OTAs that are capable of driving large capacitive loads.

However, the above FoMs do not account for the area efficiency of the OTA; in particular, no mention of the total size of the required on-chip compensation capacitance is given. For this reason, the authors of [28] introduced FoMs for large-capacitive-load OTAs:

\[
\begin{align*}
\text{LC–FOM}_{S} &= \frac{GBW}{\text{Power}} \cdot \frac{C_L}{C_t} \quad \text{LC–FOM}_{L} = \frac{SR}{\text{Power}} \cdot \frac{C_L}{C_t}
\end{align*}
\]

(12)

where \(C_t\) is the total compensation capacitance.
These FoMs give a better picture of area efficiency but become irrelevant (in the sense that they become infinite) for OTAs that do not rely on any compensation capacitors such as those in [8,49–51], which have to rely on the old FoMs in this case.

In addition, none of the above FoMs gives an indication of how complicated the compensation strategy for each architecture is: they are all performance-oriented and are calculated based on experimental results. In that sense, the same OTA architecture can have different FoM values depending on how its design was approached. These FoMs therefore do not allow for an a priori comparison between different architectures for a particular application.

At least two FoMs have been proposed that quantify the power efficiency of the compensation strategy of an OTA. The first one is the transconductance efficiency factor defined in [37] as a ratio of the OTA’s GBW to the GBW of the single stage amplifier composed of the OTA’s last stage:

$$T_e = \frac{\text{GBW} \cdot C_L}{g_{mL}} \quad (13)$$

where $g_{mL}$ is the transconductance of the last stage. The authors of [37] related the transconductance of the last stage to the GBW by imposing the Butterworth condition on the unity-gain feedback OTA while neglecting any zeros in the open-loop response. As such, the transconductance efficiency is valid for comparing different kinds of OTAs, including ones that do not employ any compensation capacitors but fail to capture the effects of open-loop zeros on the compensation strategy.

The second FoM, as suggested by the authors of [54], can be analytically evaluated in order to facilitate the comparison of compensation architectures:

$$\text{FoM}_{\text{analytical}} = \frac{\text{GBW} \cdot C_L}{g_{mL}} \quad (14)$$

where $g_{mL}$ is the sum of all transconductance values in the amplifier that require a dedicated bias current. The authors demonstrated that this FoM can be used to compare different compensation architectures analytically but again, the approach taken by the authors neglects the effects of zeros.

Finally, the authors of [59] derived a FoM that can be used to compare amplifiers with optimized settling time:

$$\text{FoM}_t = \frac{\ln \psi}{t_{\text{min}}} \frac{C_L}{g_{mL}} \quad (15)$$

where $\psi$ is the required settling accuracy in the time domain and $t_{\text{min}}$ is the minimum achievable settling time of the amplifier to an accuracy of $\psi$. This FoM is very useful for comparing architectures based on their achievable time domain performance but cannot be derived analytically since $t_{\text{min}}$ has no known closed-form solution.

For these reasons, the next section proposes a new FoM that allows for the comparison of different OTA architectures based on the power efficiency of their compensation schemes.

2.3. Proposed FoM for Architecture Selection

The Routh–Hurwitz stability criterion for a third-order polynomial $a_3s^3 + a_2s^2 + a_1s + a_0$ is $a_1a_2 - a_0a_3 > 0$. When $a_1a_2 - a_0a_3 = 0$, this indicates a pole that is about to cross into the right half-plane (RHP) and thus indicates the marginal stability of the system being studied. Since the roots of a polynomial depend in a continuous manner on its coefficients, it is reasonable to expect that as the quantity $a_1a_2 - a_0a_3$ increases, the roots of the polynomial move further into the left half-plane (LHP) rendering the system “more stable” in some sense.
Based on the above discussion, the proposed FoM for compensation efficiency is obtained by considering the OTA in unity-gain feedback without neglecting its zeros as has been done in [22,25,37]. With the coefficients \( a_3 - a_0 \) representing the closed-loop pole polynomial thus formed, the proposed FoM is defined as

\[
\kappa = \frac{a_1 a_2}{a_0 a_3}
\]  

To give an indication of how useful this FoM can be we return to the all-pole OTA whose open-loop gain is given by Equation (1). If we consider the Bode plot of the open-loop gain, sketched conceptually in Figure 5, we note that the complex pole pair contributes \(-90^\circ\) of phase shift at \( \omega = \omega_0 \). It is thus reasonable to estimate the phase crossover frequency as \( \omega_\pi \simeq \omega_0 \) since the dominant pole also contributes \(-90^\circ\) of phase shift at \( \omega_\pi \) (where it is assumed that \( \omega_\pi \gg \omega_{pd} \)). It can thus be shown that in this case, the gain margin is given by

\[
\gamma_m \simeq \frac{\omega_0}{Q}
\]

Thus, according to Equation (4), \( \kappa \) corresponds exactly to the gain margin in this case so that a higher \( \kappa \) corresponds to a higher gain margin for this OTA.

The factor \( \kappa \) can be used as a measure of how power-efficient a compensation scheme is. In particular, higher values of \( \kappa \) mean that, for the same increase in power consumption, the improvement in stability is larger. This can be demonstrated by observing the effects of doubling the value of \( \omega_0 \) (doubling the power consumption) at two different values of \( \kappa \). Figure 6 demonstrates the unity-gain feedback step response in four cases: \( \omega_0 \) is doubled from 2 to 4 at \( \kappa = 2 \) and \( \kappa = 6 \) and Table 4 summarizes the settling time in each case as well as the improvement obtained from doubling the power consumption. The results demonstrate that an increased power consumption achieves a better stability improvement for larger values of \( \kappa \).

**Table 4.** Improvement in settling time achieved by doubling the value of \( \bar{\omega}_0 \) at different values of \( \kappa \).

<table>
<thead>
<tr>
<th>( \bar{\omega}_0 )</th>
<th>2</th>
<th>4</th>
<th>( T_s / T_{s_1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \kappa = 2 )</td>
<td>1.59 µs</td>
<td>0.87 µs</td>
<td>1.82</td>
</tr>
<tr>
<td>( \kappa = 6 )</td>
<td>2.06 µs</td>
<td>0.35 µs</td>
<td>5.97</td>
</tr>
</tbody>
</table>
Figure 6. Unity-gain feedback step responses at different values of $\kappa$ and $\omega_0$. Higher $\kappa$ makes the compensation more power-efficient.

This FoM can also be used for comparing two different compensation architectures. To demonstrate this, consider another three-stage OTA with a zero in the LHP such that its open-loop gain is given by

$$A_v(s) = \frac{A_0\left(1 + \frac{s}{\omega_z}\right)}{\left(1 + \frac{s}{\omega_{pd}}\right)\left(1 + \frac{s}{\omega_0Q} + \frac{s^2}{\omega_0^2}\right)} \quad (18)$$

In this case, the closed-loop pole polynomial coefficients, assuming unity-gain feedback, are given by

$$a_0 = A_0 + 1 \simeq A_0$$

$$a_1 = \frac{A_0}{\omega_z} + \frac{1}{\omega_{pd}} + \frac{1}{\omega_0Q} \simeq \frac{A_0}{\omega_z} + \frac{1}{\omega_{pd}}$$

$$a_2 = \frac{1}{\omega_{pd}\omega_0Q} + \frac{1}{\omega_0^2} \simeq \frac{1}{\omega_{pd}\omega_0^2}$$

$$a_3 = \frac{1}{\omega_{pd}\omega_0^2}$$ \quad (19)

so that $\kappa$ is given by

$$\kappa \simeq \frac{A_0}{\omega_0Q} \cdot \left(1 + \frac{\text{GBW}}{\omega_z}\right) \quad (20)$$

This result is quite intuitively satisfying since it shows that this new OTA has a higher $\kappa$ than the one given by Equation (1) and that this improvement is due to the LHP zero. Moreover, it can be seen that this improvement becomes insignificant when the zero frequency is much larger than the GBW product of the OTA.

The main limitation of using $\kappa$ as a FoM is that it cannot deal with pole-zero cancellations. This is not a very serious limitation, however, since OTAs that rely on pole-zero cancellation usually achieve this cancellation at a specific value of the load capacitor. Any change in the load capacitor makes the OTA return to being a three-pole system and justifies the use of $\kappa$ [64].

It should also be noted that OTAs having more than three poles due to additional parasitic poles have a different Routh–Hurwitz stability condition to the one derived here. This is ignored in our work on the basis that the parasitic poles are at such high frequencies that they do not significantly affect the OTA’s stability [33,38,48]. Figure 7 summarizes the design procedure discussed so far, starting from a specified GBW, maximum load capacitance, settling time and overshoot.
Select OTA architecture based on $\kappa$

Generate curves for settling time and overshoot

Select values for $\omega_0$ and $Q$ based on the curves

Circuit implementation

Simulate

Specifications met?

Refine estimate of parasitics

End

Select $g_{m1}$ based on noise or offset requirements [57]

Calculate Miller capacitance value

No

Yes

Figure 7. Flowchart for the design procedure discussed in this work.

3. Classification of Multistage Amplifier Architectures

This section proposes a classification scheme for the existing compensation techniques of three-stage OTAs and outlines the common features in each technique. After the broad divisions are discussed briefly, each technique is examined via several example architectures with $\kappa$ used as a FoM to compare them.

The classification scheme depends on whether the inner amplifier (as defined in [54]) is Miller compensated. Specifically, let the transfer function of the two-stage inner amplifier be given by $H(s)$, as shown in Figure 8. In this case, the transfer function of the whole amplifier is given by

$$A_v(s) = -\frac{g_{m1}H(s)}{g_{o1} + sC_M \left[1 + H(s)\right]}$$

(21)

For a three-stage OTA, $H(s)$ will be second-order and may be expressed generically as

$$H(s) = \frac{A_2A_3}{1 + \frac{s}{\omega_{0i}Q_i} + \frac{s^2}{\omega_{0i}^2}}$$

(22)

where $A_2$ and $A_3$ represent the voltage gains of the second and third stages, respectively, and $\omega_{0i}$ and $Q_i$ represent, respectively, the natural frequency and quality factor of the poles of the inner amplifier.
Substituting Equation (22) into Equation (21), the following relations are found for \( \omega_0 \) and \( Q \) of the whole amplifier:

\[
\omega_0 = \sqrt{A_2 A_3} \omega_{0i} \quad Q = \sqrt{A_2 A_3} Q_i
\]  

(23)

Using the example of the NGCC amplifier in Figure 1 and neglecting \( g_{m1} \) and \( g_{m2} \), one finds

\[
H(s) \approx g_{m2} g_{m3} \left( 1 - \frac{s C_M}{g_{m3}} \right) \left( 1 + \frac{s C_M}{g_{m2} g_3} + \frac{s^2 C_1 C_M}{g_{m2} g_3} \right)
\]

so that we have

\[
\omega_{0i} = \sqrt{g_{m2} g_{m3} / (C_M C_L)} \quad Q_i = \sqrt{g_{m2} g_{m3} / g_{m3} \cdot \sqrt{C_2 / C_L}}
\]

(24)

This demonstrates that, when the inner amplifier is Miller-compensated, the inner Miller capacitor controls the quality factor of the non-dominant poles and that decreasing the load capacitance reduces the quality factor.

On the other hand, if \( C_M \) were not present in Figure 1, one can show that in this case

\[
\omega_{0i} = \sqrt{g_{m2} g_{m3} / (C_2 C_L)} \quad Q_i = \sqrt{g_{m2} g_{m3} / g_{m3} \cdot \sqrt{C_2 / C_L}}
\]

This means that, when the inner amplifier is not Miller compensated, the quality factor of the non-dominant poles depends on parasitic capacitors and increases as the load capacitor is decreased. This latter feature is responsible for increased gain peaking (and reduced gain margin) as the load capacitor is decreased in such architectures. This is because, at reduced loads the inner amplifier becomes unstable [42].

Figure 9 shows the classification scheme outlined above along with example architectures in each category. All architecture acronyms, along with citations of reference works are explained in Table 5. Note from the figure that architectures where the inner amplifier is not Miller-compensated are sub-divided into reverse nested Miller architectures which rely on a feedback capacitor to control the quality factor of the inner amplifier poles and shunt \( Q \) control architectures which use a shunt equivalent impedance to control the \( Q \) factor. It should also be noted that some compensation architectures (such as DACFC [28]) may fit into multiple categories while some architectures use techniques other than Miller compensation.

3.1. OTAs with Miller-Compensated Inner Amplifier

Figure 10 shows the compensation architectures that are studied in this section. The unifying features of this architecture type are discussed and the power efficiency of the different architectures is compared. Some typical applications that can best leverage these features are briefly touched upon as well.
Table 5. Explanation and citations for the architecture acronyms in Figures 9–11.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Acronym Explanation</th>
<th>Reference Work(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMC</td>
<td>Nested Miller Compensation</td>
<td>[66]</td>
</tr>
<tr>
<td>NMCFNR</td>
<td>Nested Miller Compensation with Feedforward and Nulling Resistor</td>
<td>[22]</td>
</tr>
<tr>
<td>DPZC</td>
<td>Double Pole-Zero Cancellation</td>
<td>[2]</td>
</tr>
<tr>
<td>AFFC</td>
<td>Active Feedback Frequency Compensation</td>
<td>[24]</td>
</tr>
<tr>
<td>TCFC</td>
<td>Transconductance with Capacitances Feedback Compensation</td>
<td>[27]</td>
</tr>
<tr>
<td>NGCC</td>
<td>Nested Gm-C Compensation</td>
<td>[7]</td>
</tr>
<tr>
<td>NGRNMC</td>
<td>Nested feed-forward Gm-stage and nulling Resistor plus Nested Miller Compensation</td>
<td>[23]</td>
</tr>
<tr>
<td>DACFC</td>
<td>Dual Active Capacitive Feedback Compensation</td>
<td>[28]</td>
</tr>
<tr>
<td>RNMCFNFR</td>
<td>Reverse NMCFNR</td>
<td>[31]</td>
</tr>
<tr>
<td>RAFFC</td>
<td>Reverse Active Feedback Frequency Compensation</td>
<td>[31]</td>
</tr>
<tr>
<td>CLQC</td>
<td>Cascode Miller Compensation with Local Q-factor Control</td>
<td>[3]</td>
</tr>
<tr>
<td>DFCFC</td>
<td>Damping Factor Control Frequency Compensation</td>
<td>[34]</td>
</tr>
<tr>
<td>IAC</td>
<td>Impedance Adapting Compensation</td>
<td>[37]</td>
</tr>
<tr>
<td>AZC</td>
<td>Active Zero Compensation</td>
<td>[11]</td>
</tr>
<tr>
<td>Feedforward Compensation</td>
<td>N/A</td>
<td>[8]</td>
</tr>
<tr>
<td>Pseudo-Single-Stage</td>
<td>N/A</td>
<td>[51,52]</td>
</tr>
<tr>
<td>Current-Mirror-Based</td>
<td>N/A</td>
<td>[49,50]</td>
</tr>
<tr>
<td>RNMC-VBNR</td>
<td>Reverse Nested Miller Compensation with Voltage Buffer and Nulling Resistor</td>
<td>[32]</td>
</tr>
<tr>
<td>AFCRFC</td>
<td>Active Feedback and Current Reuse Feedforward Compensation</td>
<td>[13]</td>
</tr>
<tr>
<td>CLIA</td>
<td>Cascade-compensated Local Impedance Attenuation</td>
<td>[41]</td>
</tr>
<tr>
<td>ULLF</td>
<td>Upper-Load-Limit-Free (name given by the authors of this work for easy reference)</td>
<td>[46]</td>
</tr>
</tbody>
</table>
Three-Stage OTAs

Miller-Compensated Inner Amplifier

Uncompensated Inner Amplifier

Other Techniques

Feedforward Compensation

Reverse NM

Shunt Q Control

Pseudo-Single-Stage Amplifiers

Gain Booster

Current-Mirror-Based Amplifiers

m Amplifying Current Mirrors

Figure 9. Classification of three-stage OTA compensation techniques with example architectures given for each category. Cyan blocks indicate parts of the core amplifier while magenta blocks indicate components used for compensation.
As expected, all architectures (with the exception of DACFC) share the feature that $Q$ decreases as the load capacitance is decreased. DACFC is a special case because both compensation loops share the same capacitor, thus it does not fit the structure of Figure 8.

Table 6 shows expressions for $\kappa$, $\omega_0$ and $Q$ for the compensation architectures of Figure 10. The abbreviated architecture names are explained in Table 5.
Table 6. Analytical expressions for the proposed FoM and non-dominant pole parameters for the compensation architectures of Figure 10.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>$\kappa$</th>
<th>$\omega_0$</th>
<th>$Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMC</td>
<td>$\frac{C_{M1}}{C_L} \left( \frac{g_{m1} - g_{m2} - g_{m3}}{g_{m3}} \right)$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_{M1} C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{g_{m1} - g_{m2}}}$</td>
</tr>
<tr>
<td>NGCC</td>
<td>$\frac{C_{M1}}{C_L} \left( \frac{g_{m1} + g_{m2} (g_{m3} R_{M1} - 1) + g_{m2}}{g_{m3}} \right) \left[ 1 + \frac{C_{M2}}{C_{M1}} \right]$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_{M1} C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{g_{m1} + g_{m2} - g_{m3}}}$</td>
</tr>
<tr>
<td>NMCFCN</td>
<td>$\frac{C_{M1}}{C_L} \left( \frac{g_{m1} + g_{m2} (g_{m3} R_{M1} - 1)}{g_{m3}} \right) \left[ 1 + \frac{g_{m1} R_{M1}}{g_{m3}} \right]$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_{M1} C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{g_{m1} + g_{m2} - g_{m3}}}$</td>
</tr>
<tr>
<td>NGRNMC</td>
<td>$\frac{C_{M1}}{C_L} \left( \frac{g_{m1} + g_{m2} (g_{m3} R_{M1} - 1)}{g_{m3}} \right) \left[ 1 + \frac{g_{m1} R_{M1}}{g_{m3}} \right]$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_{M1} C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{g_{m1} + g_{m2} - g_{m3}}}$</td>
</tr>
<tr>
<td>DPZC</td>
<td>$\frac{C_{M1}}{C_L} \left( \frac{1 + g_{m2} R_{M2} g_{m3} R_{M1}}{g_{m3} R_{M1}} - 1 \right) \left[ \frac{g_{m1} + g_{m2} (1 + g_{m2} R_{M2} g_{m3} R_{M1}) - g_{m2}}{g_{m1}} \right] \left[ 1 + \frac{g_{m1} R_{M1}}{g_{m3}} \right] + R_{M2} \left( \frac{1 + g_{m1} R_{M1}}{g_{m3}} \right)$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_{M1} C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{g_{m1} + g_{m2} R_{M2} - g_{m2}}}$</td>
</tr>
<tr>
<td>DACFC</td>
<td>$\frac{C_{M1}}{C_L} \left( \frac{g_{m1} - g_{m2} - g_{m3}}{g_{m3}} \right)$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_{M1} C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{g_{m1} - g_{m2}}}$</td>
</tr>
<tr>
<td>TCFC</td>
<td>$\frac{C_{M1}}{C_L} \left( \frac{g_{m1} - g_{m2} - g_{m3}}{g_{m3}} \right) \left( \frac{1 + g_{m1} R_{M1}}{g_{m3}} \right)$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_{M1} C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{g_{m1} + g_{m2} - g_{m3}}}$</td>
</tr>
<tr>
<td>AFFC</td>
<td>$\frac{C_{M1}}{C_L} \left( \frac{g_{m1} - g_{m2}}{g_{m3}} \right)$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{C_{M1} C_L}}$</td>
<td>$\sqrt{\frac{g_{m2} g_{m3}}{g_{m1} - g_{m2}}}$</td>
</tr>
</tbody>
</table>
Another interesting feature to note is that architectures that feature some form of cascode compensation (the bottom group in the table) have their $\kappa$ values enhanced by a factor that is the ratio of a Miller capacitor to a parasitic capacitor. This means that these compensation architectures tend to be more power-efficient as evidenced also by their $\omega_0$ expressions being enhanced by similar factors with respect to their counterpart architectures (the top group in the table). This enhancement, observed in two-stage amplifiers as well, is due to one or both of the amplifier’s internal nodes avoiding the loading effect from the Miller capacitor (see, for example, [11,67]).

A common feature to these architectures is their reduced power efficiency at large capacitive loads. This may be seen both from the $\kappa$ and the $\omega_0$ expressions. Large values of $\kappa$ or $\omega_0$ (increasing $\omega_0$ means increasing $\sqrt{\omega_0}$ and is desirable for improved settling time) require an increased power consumption.

The situation becomes worse due to the feedforward current through the inner Miller capacitor. This current creates a RHP zero in $H(s)$ and, by reference to Equation (21) adds negative terms to the amplifier’s transfer function denominator requiring an increased value of $g_m3$ to avoid the amplifier going unstable (for example, to avoid $\kappa$ and $Q$ going negative in the case of NMC). Intuitively, it means that the current through $g_m3$ needs to be made a lot larger than the feedforward current through $C_m2$. This translates into the $-g_m2$ terms observed in the $\kappa$ factors of NMC and DPZC for instance. DPZC may overcome this problem by properly sizing the nulling resistor $R_m2$ to push this RHP zero to the LHP. Again, as may be observed from the $\kappa$ expressions, architectures that use cascode compensation to block the RHP zero through the inner Miller capacitor (DACFC and TCFC) or feedforward paths to cancel it (NGCC) do not suffer from this problem and are therefore more power-efficient.

As stated above, the $\kappa$ expressions allow us to compare architectures prior to designing them. With reference to Table 6, the power efficiency advantage of NGCC over NMC is immediately obvious (in the sense that $\kappa$ is larger for the same $g_m3$ value). It is also observed that NMCFNR, NGRNMC and DPZC all have better $\kappa$ than NGCC for the same power consumption at the expense of added complexity. It should be noted be noted that the presence of resistive terms in these $\kappa$ expressions means that the designer can trade off chip area (and noise performance) to achieve better power efficiency by increasing the relevant resistor sizes. More subtly, it can be seen that NGRNMC allows is slightly more efficient than DPZC in that sense since its resistive terms are amplified by $g_m3$ which tends to be larger than the $g_m2$ terms seen in DPZC.

Moving towards the bottom group of the table, we observe the aforementioned enhancement to $\kappa$ over the top group. Note that this enhancement is inherent to the architecture and does not require an extra power expenditure (the currentbuffers in cascode compensation are often embedded in the first and second stages and therefore exploit current reuse to achieve improved power efficiency at no additional power cost). More importantly, comparing $\kappa$ expressions, it is evident that using cascode compensation in the inner loop leads to more power efficiency than using it in the outer loop.

Finally, note that DACFC is a special case in the sense that its $\kappa$ value is independent of the load capacitance and should therefore have superior power efficiency to the other architectures in this category.

Applications that best leverage the properties of this type of architecture are applications where the maximum load capacitance is limited but where the amplifier should be stable at reduced load capacitance. A typical example of such an application are output-capacitor-free low-dropout (LDO) linear regulators that drive an on-chip power network with limited capacitance (see, for example, [68]). Such applications fit with this type of architecture since $g_m3$ will be the transconductance of the LDO power transistor and will therefore be large by design. If settling speed and/or power efficiency are really critical, architectures with improved $\kappa$ such as DACFC or TCFC should be preferred.

3.2. OTAs with Uncompensated Inner Amplifier

Figure 11 shows the compensation architectures that will be studied in this section. The discussion follows the same order as the previous section.
Figure 11. Compensation architectures with uncompensated inner amplifiers. Cyan blocks show the core amplifier stages while magenta blocks show additional transconductors used for compensation.

Table 7 shows expressions for $\kappa$, $\omega_0$ and $Q$ for the compensation architectures of Figure 11.

First note that all architectures share the feature that $Q$ increases as the load capacitance is decreased. This leads to increased gain peaking (and decreased gain margin) when the load is decreased as discussed above.
Table 7. Analytical expressions for the proposed FoM and non-dominant pole parameters for the compensation architectures of Figure 11.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>$\kappa$</th>
<th>$\omega_0$</th>
<th>$Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAFFC</td>
<td>$\frac{g_m}{g_m}$ (1 + \frac{g_m}{g_m} )</td>
<td>$\sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\frac{C_{M1}}{C_{M2}C_L}$</td>
</tr>
<tr>
<td>RNMC-VBNR</td>
<td>$\frac{g_m}{g_m} \left[ 1 + g_m \left( R_M + R_M + r_v C_{M1} \right) \right]$</td>
<td>$\sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\frac{C_{M1}}{C_{M2}C_L}$</td>
</tr>
<tr>
<td>AFCRFC</td>
<td>$\frac{g_m}{g_m} \left( 1 + \frac{g_m}{g_m} + \frac{C_{M2}C_m}{C_m} \cdot \frac{g_m}{g_m} \right)$</td>
<td>$\sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\frac{C_{M1}}{C_{M2}C_L}$</td>
</tr>
<tr>
<td>DFCFC</td>
<td>$\frac{C_{M1}}{C_m} \cdot \frac{g_m}{g_m} \cdot \left( 1 - \frac{g_m}{g_m} \right) \sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\frac{C_{M1}}{C_{M2}C_L}$</td>
</tr>
<tr>
<td>IAC</td>
<td>$\frac{C_m}{C_m} \cdot \left( 1 + \frac{g_m}{g_m} \right)$</td>
<td>$\sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\frac{C_{M1}}{C_{M2}C_L}$</td>
</tr>
<tr>
<td>AZC</td>
<td>$\frac{C_m}{C_m} \cdot \left( 1 + \frac{g_m}{g_m} \right)$</td>
<td>$\sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\frac{C_{M1}}{C_{M2}C_L}$</td>
</tr>
<tr>
<td>CLIA</td>
<td>$\frac{g_m}{g_m} \cdot \left( 1 + \frac{g_m}{g_m} \right)$</td>
<td>$\sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\frac{C_{M1}}{C_{M2}C_L}$</td>
</tr>
<tr>
<td>ULLF</td>
<td>$\frac{g_m}{g_m} \left( g_m C_m + g_m C_L \right)$</td>
<td>$\sqrt{\frac{g_m}{g_m}} \cdot \sqrt{C_{M1}C_L}$</td>
<td>$\frac{C_{M1}}{C_{M2}C_L}$</td>
</tr>
</tbody>
</table>
Note that the architectures discussed here are divided into the reverse nested Miller type and the shunt $Q$ control type. Amplifiers of the first type have their $\omega_0$ values depend on the $Q$ control capacitor as it loads the output of the second stage. When shunt $Q$ control is used, there is no capacitor loading the output of the second stage, which leads to much higher values of $\omega_0$ for the same power consumption. This is the main reason that such architectures are well-suited to drive ultra large capacitive loads (in the nF range), as seen in recent publications [11,41,42,45].

Another feature common to all these architectures is that $\kappa$ is independent of the load capacitor, which means that the amplifier retains its power efficiency over a wide range of load capacitance. For this reason, these architectures are well-suited to drive a wide range of capacitive loads and can even have no upper limit on the capacitive load they can drive [46].

On the other hand, the increased gain peaking at reduced load sets a lower limit on load capacitance. It should be noted that a strictly three-pole OTA from these architectures that has a gain margin defined by Equation (17) will have a load-capacitance-independent gain margin and therefore no lower limit on load capacitance. In practice, however, this is not the case as practical OTAs have parasitic poles. To illustrate this, define $\omega_{p4}$ as the frequency of a parasitic pole such that the open loop transfer function of the OTA may be expressed as

$$A_v(s) = \frac{A_0}{\left(1 + \frac{s}{\omega_{p4}}\right) \left(1 + \frac{s}{\omega_0 Q} + \frac{s^2}{\omega_0^2}\right) \left(1 + \frac{s}{\omega_{p4}}\right)}$$ (27)

If the Routh–Hurwitz criterion is applied to the open-loop transfer function, the following condition is obtained

$$\frac{\omega_{p4}}{\omega_0 Q} > 1$$ (28)

All the architectures in Table 7 have $\omega_0 Q \propto C_L$. Let $\omega_0 Q = a C_L$ for some constant $a$. This transforms Equation (28) into the condition

$$C_L > \frac{a}{\omega_{p4}}$$ (29)

so that minimizing the lower limit on the load capacitance requires pushing the parasitic pole to a high frequency and having a low value for the constant $a$. This fact demonstrates the conflicting requirements for driving a wide range of load capacitance. Designing the amplifier for a given phase margin or settling time at a large value of $C_L$ requires a large value of $\omega_0$ and therefore a large value of $a$. Thus, pushing the upper limit on $C_L$ automatically increases the lower limit as well and limits the overall range of loads the OTA can drive.

As before, the $\kappa$ expressions allow us to compare architectures and we see that the shunt $Q$ control types tend to have larger values of $\kappa$ due to their inverse dependence on parasitic capacitors. Note also that the expressions allow us to see equivalences among architectures as well. By comparing the expressions for DFCFC and IAC, for instance, one sees that their efficiency is roughly equivalent when $g_{m4} = \frac{1}{R_a}$, which makes sense when their block diagrams are compared. This shows that the same performance can be obtained from either architecture and the designer should make the choice based on whether a larger area ($R_a$) or power ($g_{m4}$) expenditure can be tolerated.

This type of architecture is best suited for applications that require driving a very wide range of load capacitors such as headphone drivers [15,46], LCD drivers [11,14], capacitive MEMS sensors [14,46], and LDO regulators [46]. For LDO regulators, however, care should be taken that $C_2$ is the gate capacitance of the power transistor and can therefore be substantially large.

3.3. Other Compensation Architectures

Some architectures do not use any Miller capacitors to achieve stability and are briefly surveyed here for the sake of completeness.
Feedforward compensation [8] uses feedforward transconductance stages to generate zeros whose phase shift compensates that of the OTA poles thereby achieving stability. This architecture requires large power consumption in the feedforward stages to pull the zeros to low enough frequencies to be close to the poles.

The amplifiers in [51,52] achieve pseudo single-stage behavior by reducing the impedance of inner amplifier nodes to push the non-dominant poles to high frequencies. The required DC gain is restored by using shunt gain booster circuits, as shown in Figure 9.

As an alternative approach, the amplifiers in [49,50] convert the input voltage into current then rely on cascading current amplification stages (amplifying current mirrors) to achieve the required DC gain. This approach leads to superior power efficiency when driving ultra large load capacitors, but the parasitic poles created by the cascade of current mirrors limit the load driving range as explained above.

4. Circuit-Level Considerations

The aim of this section is to bring attention to two issues that pertain to the circuit-level realization of the OTA architectures discussed above. One issue relates to using cascode compensation and the other to using the OTA to drive ultra large load capacitors.

4.1. Cascode Compensation

OTAs that implement cascode compensation often eschew implementing a dedicated current buffer by reusing the current in the input stage, usually realized as a folded cascode [11,33,42,45,46,48]. Some architectures use a traditional input stage with a current mirror load and use its low-impedance node as an embedded capacitive amplifier [39,69]. While these approaches lead to superior power efficiency at large capacitive loads, they tend to add parasitic poles to the inner amplifier’s transfer function rendering its stability quite sensitive to the load capacitor and severely limiting the drivable load range.

Designers should be aware that this problem is exacerbated by the desire to operate the amplifier at low power since this limits the impedance at the low-impedance nodes of the folded cascode and current mirror stages. In such architectures, there is a clear tradeoff between power consumption and drivable load range.

Another factor responsible for introducing parasitic poles into the inner amplifier transfer function is the restriction that either stage $g_m_2$ or $g_m_3$ must be non-inverting, as may be observed from Figures 10 and 11. This introduces at least one additional parasitic pole since a current mirror is needed to invert the polarity of a class-A stage. One way to solve this is to connect the Miller capacitor into the correct arm of the folded cascode in order to implement an inverting current buffer and allow both $g_m_2$ and $g_m_3$ to be inverting as explained in [46].

4.2. Slew Rate Enhancers

Amplifiers that drive ultra-large load capacitors tend to have their slew rate limited by the load capacitor. This situation is undesirable if the output stage is of class-A type since either the charging or discharging current of the load capacitor will be limited, which leads to a severely degraded rise or fall times.

Many architectures solve this by connecting a feedforward transconductance $g_m_f$ to the output node, as seen in Figures 10 and 11. This makes the output stage behave in a push-pull manner similar to a class-AB stage and improves the slew rate significantly.

Another alternative is to implement dedicated slew rate enhancers (SREs). SREs are circuits that draw no current during normal operation but turn on when the amplifier is slewing and provide extra charging and/or discharging current to the load capacitor as needed. Several previous works have discussed these circuits [14,32,46,70].
5. Confirmation of Results through Transistor Simulations

To further validate the above discussion, representative architectures are selected from the ones discussed above and are implemented at the transistor level in a standard 0.18 µm CMOS technology and their performance is compared via simulations using BSIM4, level 14 MOSFET models with the Spectre simulator®. For a fair comparison, all OTAs are implemented with the same power budget whenever possible. The selected architectures for the comparison are NMC, NMCFNR, TCFC, RAFFC, IAC and CLIA.

5.1. Schematic Diagrams

To achieve a fixed power budget, all OTA implementations share the same core circuit with the exception of OTAs of the reverse nested Miller type, which require a non-inverting third stage. The circuit schematics for both types of OTA cores are shown in Figure 12. Note that the magenta devices represent auxiliary devices for compensation. M_{fb} is used as a feedback current buffer for architectures that employ cascode compensation. M_{buff} serves a similar purpose for TCFC only and is removed in other architectures. Finally, the M_{ff} devices implement a feedforward transconductance and, accordingly, V_x may be connected to the output of the third stage or to a bias voltage depending on whether the OTA has a feedforward transconductance or not.

Figure 12. Schematic diagrams for the two OTA cores used in simulations. Magenta transistors indicate auxiliary transconductance elements used for compensation. M_{buff} in (a) is used as a current buffer in TCFC only and is removed in other architectures.

In addition to the amplifier cores, the SRE described in [46] is implemented to add additional charging/discharging current at the output node of Figure 12a,b, respectively. Tables 8 and 9 show the device sizes for the devices in Figure 12a,b, respectively.
Table 8. Device sizes for the transistors in Figure 12a.

<table>
<thead>
<tr>
<th>Device</th>
<th>W/L (µm/µm)</th>
<th>Device</th>
<th>W/L (µm/µm)</th>
<th>Device</th>
<th>W/L (µm/µm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M11</td>
<td>1.6/0.6</td>
<td>M18</td>
<td>2.4/0.6</td>
<td>M_{buff}</td>
<td>2.4/0.6</td>
</tr>
<tr>
<td>M12</td>
<td>0.7/0.18</td>
<td>M_{fb}</td>
<td>2.4/0.6</td>
<td>M18</td>
<td>1.6/0.6</td>
</tr>
<tr>
<td>M13</td>
<td>0.7/0.18</td>
<td>M_{10}</td>
<td>0.8/0.18</td>
<td>M_{15}</td>
<td>57.6/0.6</td>
</tr>
<tr>
<td>M14</td>
<td>2.4/0.6</td>
<td>M_{11}</td>
<td>0.8/0.6</td>
<td>M22</td>
<td>16/0.6</td>
</tr>
<tr>
<td>M15</td>
<td>2.4/0.6</td>
<td>M_{21}</td>
<td>4.8/0.6</td>
<td>M_{13}</td>
<td>24/0.18</td>
</tr>
<tr>
<td>M16</td>
<td>0.8/0.18</td>
<td>M_{22}</td>
<td>0.8/0.6</td>
<td>M_{23}</td>
<td>4.8/0.6</td>
</tr>
<tr>
<td>M17</td>
<td>0.8/0.6</td>
<td>M_{23}</td>
<td>6.4/0.6</td>
<td>M_{13}</td>
<td>0.8/0.6</td>
</tr>
</tbody>
</table>

Table 9. Device sizes for the transistors in Figure 12b.

<table>
<thead>
<tr>
<th>Device</th>
<th>W/L (µm/µm)</th>
<th>Device</th>
<th>W/L (µm/µm)</th>
<th>Device</th>
<th>W/L (µm/µm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M11</td>
<td>1.6/0.6</td>
<td>M18</td>
<td>2.4/0.6</td>
<td>M32</td>
<td>1/0.6</td>
</tr>
<tr>
<td>M12</td>
<td>0.7/0.18</td>
<td>M_{fb}</td>
<td>2.4/0.6</td>
<td>M18</td>
<td>64/0.6</td>
</tr>
<tr>
<td>M13</td>
<td>0.7/0.18</td>
<td>M_{10}</td>
<td>0.8/0.18</td>
<td>M33</td>
<td>6.4/0.6</td>
</tr>
<tr>
<td>M14</td>
<td>2.4/0.6</td>
<td>M_{11}</td>
<td>0.8/0.6</td>
<td>M_{34}</td>
<td>24/0.6</td>
</tr>
<tr>
<td>M15</td>
<td>2.4/0.6</td>
<td>M_{21}</td>
<td>6.4/0.6</td>
<td>M_{2}</td>
<td>0.8/0.6</td>
</tr>
<tr>
<td>M16</td>
<td>0.8/0.18</td>
<td>M_{22}</td>
<td>1.6/0.6</td>
<td>M_{3}</td>
<td>1.6/0.6</td>
</tr>
<tr>
<td>M17</td>
<td>0.8/0.6</td>
<td>M_{31}</td>
<td>11.2/0.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.2. Selection of Circuit Parameters

The design is approached from the time domain perspective as discussed above. A settling time of less than 2 µs with a GBW of 1 MHz is targeted. An overshoot of less than 15% is also specified. The GBW value is set as a round figure approximation to the value achieved in several recent designs [3,42,46]. With this value of GBW, the normalized settling time \( T_s \times \text{GBW} \) is closest to the value 10 used as a heuristic in [62]. Finally, an overshoot of 15% roughly corresponds to a second-order system with a damping factor of 0.5 [65].

For each architecture, normalized curves similar to those in Figure 4a,b are generated using numerical simulations in order to choose values of \( \omega_0 \) and \( Q \) that satisfy the performance targets. As an example, Figure 13 shows the curves generated for the NMC architecture and the selected design point (\( \omega_0 = 2.5, Q = 0.9 \)) where \( \omega_0 = 2.5 \) was chosen to have an overshoot below 15% over a wide range of \( Q \) values and \( Q \) was chosen to optimize overshoot and settling time.

Figure 13. Curves used for the choice of design parameters for the NMC architecture: (a) normalized settling time and (b) overshoot. The black dot marks the selected design point.

Using the selected values of \( \omega_0 \) and \( Q \) along with the expressions for \( \omega_0 \) and \( Q \) in Tables 6 and 7, component values for the compensation network are chosen after \( g_{m2} \) and \( g_{m3} \) are fixed by the common...
OTA core. Since \( g_{m2} \) and \( g_{m3} \) are fixed beforehand in these example designs, the expressions for \( \omega_0 \) and \( Q \) are instead used to calculate the load capacitance that the OTA is capable of driving while achieving the given time domain performance.

5.3. Simulation Results

For simulations, the OTA cores were implemented in a standard 0.18 \( \mu \)m CMOS technology with targeted values for \( g_{m1} \), \( g_{m2} \) and \( g_{m3} \) set to 10 \( \mu \)S, 50 \( \mu \)S and 500 \( \mu \)S, respectively. To account for routing and layout parasitics, the capacitances \( C_1 \) and \( C_2 \) were assumed to be 10 fF each (Based on extracted capacitance data from metal traces in the layout.).

For each architecture, the chosen compensation components are added and the resulting OTA is simulated to determine the range of load capacitance that each OTA can drive. The simulation results are shown in Table 10.

The maximum load capacitance is reported as the load capacitance at which either the settling time exceeds 2 \( \mu \)s or the phase margin drops below 45° so that exceeding ringing appears in the step response. The minimum load capacitance, whenever a finite number is reported, is the minimum load capacitance below which the amplifier’s open-loop transfer function has a RHP pole due to the parasitic poles as explained above.

It should be noted that the differences in DC power consumption and core \( g_{m} \) values are due to the different biasing conditions according to whether the node \( V_x \) is connected to a constant bias voltage or to the output of the first stage. In addition, note that \( g_{m6} \) of the CLIA architecture is considerably larger than the other cases because it was enhanced by a \( g_{m} - \)boosting amplifier (as explained in [11]) (the gain-boosting amplifier was implemented using an ideal voltage-controlled voltage source with a gain of 33 dB for the purpose of this simulation to avoid disturbing the bias point of the core amplifier as much as possible) because a large value is needed for it to keep the OTA \( Q \) to a low value and avoid gain peaking (as may be observed from Table 7). The value of \( g_{m3} \) had to be increased as well to achieve an acceptably large \( \omega_0 \).

The results confirm the observations made in previous sections. By comparing NMC, NMCFNR and TCFC, we note that adding the resistor improves the load driving capability only slightly while blocking the RHP zero in the inner amplifier leads to a more than 50\( \times \) improvement in the maximum drivable load capacitance.

When architectures with non-compensated inner amplifiers are used, the required auxiliary compensation capacitor decreases significantly. We note that the reverse nested Miller type amplifier does not exhibit the same efficiency as the other amplifiers as it keeps the internal node loaded with a large capacitor as explained above; in addition, one has to impose a minimum load capacitance on it for stability.

Finally, note that CLIA, a similar architecture to IAC, can achieve compensation using a much smaller physical resistor due to the added cascode compensation. The tradeoff is that a larger power consumption is needed to push the parasitic poles to high frequencies in this case since cascode compensation causes the inner amplifier to have a larger-order transfer function and to therefore be more susceptible to instability as the load capacitance is decreased.

Figure 14 shows the average of the rising and falling settling times, phase margin and gain margin as functions of the load capacitance for each architecture where it can be observed once again how a large phase margin is not necessary for good settling performance where the NMC architecture for example exhibits a settling time of 1 \( \mu \)s and an overshoot of 2\% at a phase margin of about 40°. Furthermore, note how all architectures have a phase margin that increases with decreased load while the gain margin in architectures with non-compensated inner amplifiers stays constant or gets worse with decreased load capacitance as expected due to the increased gain peaking.
Table 10. Simulation results for the selected OTA architectures.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>$I_{DD}$ (µA)</th>
<th>$A_o$ (dB)</th>
<th>$g_{m_1}$ (µS)</th>
<th>$g_{m_2}$ (µS)</th>
<th>$g_{m_3}$ (µS)</th>
<th>$g_{m_4}$ (µS)</th>
<th>$g_{m_{ff}}$ (µS)</th>
<th>$g_{m_{fb}}$ (µS)</th>
<th>$R$ (kΩ)</th>
<th>$C_{M_1}$ (pF)</th>
<th>$C_{M_2}$ or $C_z$ (pF)</th>
<th>$C_{L_{min}}$ (pF)</th>
<th>$C_{L_{max}}$ (pF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMC</td>
<td>53.5</td>
<td>112</td>
<td>10</td>
<td>67</td>
<td>499</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.5</td>
<td>4</td>
<td>0</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>NMCFNR</td>
<td>43.5</td>
<td>123.7</td>
<td>10</td>
<td>67</td>
<td>396</td>
<td>420</td>
<td>–</td>
<td>10</td>
<td>1.5</td>
<td>2.3</td>
<td>0</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>TCFC</td>
<td>40</td>
<td>127</td>
<td>10</td>
<td>63</td>
<td>367</td>
<td>394</td>
<td>38.9</td>
<td>–</td>
<td>1.5</td>
<td>4.6</td>
<td>0</td>
<td>2500</td>
<td></td>
</tr>
<tr>
<td>RAFFC</td>
<td>78.4</td>
<td>124</td>
<td>11</td>
<td>63.5</td>
<td>500.5</td>
<td>655</td>
<td>20</td>
<td>–</td>
<td>1.5</td>
<td>0.15</td>
<td>120</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>IAC</td>
<td>43.5</td>
<td>123.7</td>
<td>10</td>
<td>67</td>
<td>396</td>
<td>420</td>
<td>–</td>
<td>5700</td>
<td>1.5</td>
<td>0.15</td>
<td>0</td>
<td>1700</td>
<td></td>
</tr>
<tr>
<td>CLIA</td>
<td>80.9</td>
<td>102</td>
<td>10</td>
<td>106</td>
<td>650</td>
<td>655</td>
<td>660</td>
<td>30</td>
<td>1.5</td>
<td>0.15</td>
<td>0</td>
<td>4850</td>
<td></td>
</tr>
</tbody>
</table>
Finally, to confirm the utility of $\kappa$, Figure 15 plots the value of $\kappa$ as a function of load capacitance for each architecture. Where the superiority of the TCFC architecture is observed. Furthermore, the plot shows that, if the load capacitance is small, it is more power-efficient to use an architecture with Miller-compensated inner amplifier.

**Figure 14.** Simulation results for the chosen OTA architectures showing: (a) settling time; (b) overshoot; (c) phase margin; and (d) gain margin as functions of the load capacitance.

**Figure 15.** The value of $\kappa$ estimated from simulation results as a function of load capacitance for each architecture.
6. Conclusions

A thorough study of three-stage OTA architectures is presented, surveying a broad selection of state-of-the-art compensation techniques. The review emphasizes the point that designing these OTAs based on a target phase margin can lead to wasted power and degraded settling performance and that the main design targets need to be time domain performance parameters such as settling time and overshoot.

After discussing features common to all three-stage OTAs, a novel FoM is proposed to enable architecture selection prior to designing for a target application. In addition, a classification of the many different OTA architectures is proposed to expose the common features to each type of architecture and explain its suitability to different types of applications.

The observations presented were confirmed with transistor-level simulation results, which were found to agree with the expectations from the theoretical discussion.

Future extensions to this work should include a comparison of the linearity and noise performance of different OTA architectures from both an architecture perspective and a circuit perspective. Ideally, such an analysis should build on the work in [56] and generalize it to multiple three-stage OTA architectures to identify desirable features in the compensation architecture from a non-linearity standpoint. A similar study should be carried out for noise performance as well.

Another avenue for future work is to compare the area requirements of different architectures and extend the verification methodology by laying out the different OTA designs and comparing post-layout simulation results.

Author Contributions: Conceptualization, J.R., J.J.E.-L. and E.S.-S.; formal analysis, J.R.; data curation, J.J.E.-L.; supervision, J.J.E.-L. and E.S.-S.; writing - original draft, J.R.; writing - review & editing, J.R., J.J.E.-L. and E.S.-S.

Funding: This work was funded in part by the National Science Foundation under Grant No. ECCS-1810125.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this Appendix, the phase margin estimate of Equation (9) is derived. Start from

$$\phi_m \simeq 90^\circ + \phi_p$$  \hspace{1cm} (A1)

where $\phi_p$ is the phase shift due to the complex pole pair. This phase shift depends on the position of the UGF $\omega_\mu$ as

$$\phi_p = \begin{cases} 
- \arctan \left[ \frac{\omega_\mu}{\omega_0} \right] & \omega_\mu \leq \omega_0 \\
180^\circ - \arctan \left[ \frac{\omega_\mu}{\omega_0} \left( \frac{\omega_\mu}{\omega_0} \right)^2 - 1 \right] & \omega_\mu > \omega_0
\end{cases}$$

$$\tan (\phi_p) = -\frac{\omega_\mu}{\omega_0 Q} \quad \forall \omega_\mu$$  \hspace{1cm} (A2)
Substituting Equation (A2) into Equation (A1) yields

\[
\frac{\tan(90^\circ - \phi_m)}{\tan(\phi_p)} \simeq -\frac{\omega_p}{\omega_0 Q} \simeq -\frac{1}{1 - \left(\frac{\omega_p}{\omega_0 Q}\right)^2}
\]

\[
\phi_m \simeq \arctan\left[\frac{1 - \left(\frac{\omega_p}{\omega_0 Q}\right)^2}{\left(\frac{\omega_p}{\omega_0 Q}\right)}\right]
\]

(A3)

References

3. Cheng, Q.; Li, W.; Tang, X.; Guo, J. Design and Analysis of Three-Stage Amplifier for Driving pF-To-nF Capacitive Load Based on Local Q-Factor Control and Cascode Miller Compensation Techniques. Electronics 2019, 8, 572. [CrossRef]
11. Yan, Z.; Mak, P.L.; Law, M.K.; Martins, R.P. A 0.016-mm$^2$ 144-$\mu$W Three-Stage Amplifier Capable of Driving 1-to-15 nF Capacitive Load with >0.95-MHz GBW. IEEE J. Solid-State Circuits 2013, 48, 527–540. [CrossRef]


40. Qu, W.; Im, J.P.; Kim, H.S.; Cho, G.H. A 0.9V 6.3µW Multistage Amplifier Driving 500pF Capacitive Load with 1.34MHz GBW. In Proceedings of the 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, USA, 9–13 February 2014; doi:10.1109/isscc.2014.6757438. [CrossRef]


49. Mak, K.H.; Lau, M.W.; Guo, J.; Mui, T.W.; Ho, M.; Goh, W.L.; Leung, K.N. A 0.7 V 24 µA Hybrid OTA Driving 15 nF Capacitive Load with 1.46 MHz GBW. *IEEE J. Solid-State Circuits* 2015, 50, 2750–2757. [CrossRef]


