Modeling and Analysis of Adaptive Frequency Synthesis for Supply Droop Mitigation

Sunjin Choi, *Student Member*, *IEEE*, Hesham Beshary, *Student Member*, *IEEE*
{sunjin_choi, hesham.beshary}@berkeley.edu

***Abstract*—Dynamically changing clock frequencies to guard system against timing failure is a promising technique for power/performance optimization in modern commercial processors. Also known as Adaptive Clocking scheme, it can reduce Vdd guardband and thus the minimum achievable supply voltage Vmin while guaranteeing error-free operations. Adaptive Frequency Synthesis is one of such techniques to mitigate the effect of high-frequency supply noise at hundreds of megahertz by directly adapting the PLL-driven clock at the instant of droop detection. However, previous studies lack system-level analysis of such systems except that presented in [1]. In this paper, we present a system model for AFS systems by extending the model in [1] and accommodating a simplified PLL model in Verilog-A. Based on the model, we expect to derive the relationship between design parameters like PLL loop bandwidth, frequency step, clock insertion delay and system performance metric Vmin.**

***Index Terms*—Adaptive Frequency Synthesis (AFS), Adaptive Clocking (AC), Power Supply Noise (PSN)**

I. Introduction

Supply voltage droop imposes a critical limitation to processor power optimization while ensuring the timing constraints to be met across wide operating conditions. While supply droop transients occur infrequently during the system operation, designers should allocate a certain amount of supply margin to tolerate any worst-case supply variations. However, it is also beneficial to reduce the supply margin as reducing Vdd directly translates into power efficiency gain. Alternative approaches to reduce excess supply margin is to adaptively tune the clocking circuits during a droop event to avoid potential timing failures. One of the popular implementations is Adaptive Frequency Synthesis (AFS) which directly modulates PLL-driven clocks to compensate for the delay changes at critical paths. In this paper, we propose a model of AFS system to study the effects of various design parameters on the system performance.

A challenge with designing adaptive clocking circuits is to ensure correct functionalities under a large supply droop of a few millivolt at a few hundreds of megahertz speed. Traditionally, Vdd guardband was employed to guard against supply variations without intricate clocking schemes. Introduction of adaptive techniques reduced Vdd guardband to enhance the power efficiency of the processors. Fig. 1 illustrates such benefit. However, the requirements of adaptive clocking are getting more stringent with the processor generation. For example, it should be (1) fast enough to mitigate nanosecond-scale fast droops, (2) implemented with low area/power overhead, (3) calibrated across DVFS modes.

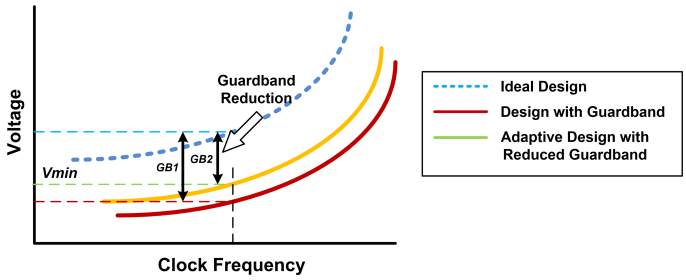
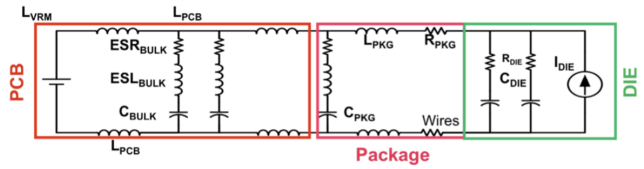


Fig. 1: Guardband Reduction with Adaptive Design

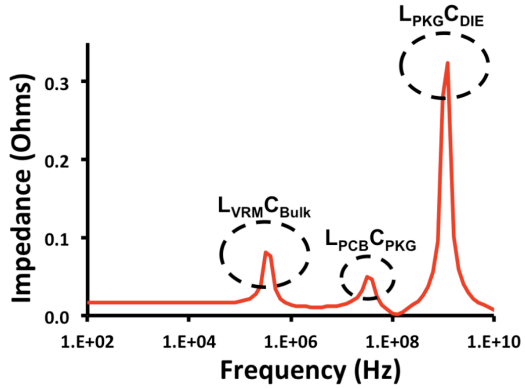
Numerous techniques have been proposed over 15 years to meet those requirements like Clock Stretcher (CS) [2]–[4], Adaptive Clock Distribution (ACD) [5], [6], Proactive Clock Gating (PCG) [7], [8], etc. Among these techniques, AFS can minimize the area/power overhead as it can be implemented with a droop detector and corresponding PLL controls.

As adaptive clocking systems in general are required to quickly adapt the clock frequency to the supply droop and the slow-down of datapath delay, it is important to model the response latency of the systems and correlate various design parameters to the delay. Response Delay Model proposed in [1] identifies the design constraints of the subsystems by modeling the response latency as a sum of detection and response delays. Their impact on the system is then investigated in terms of system performance metric Vmin, or minimum achievable supply voltage. Also a phenomena called Clock-Data Compensation (CDC) is studied in [9], [10] with a small signal delay model to analyze the effect of various design parameters and served as a basis of ACD circuit designs.

In this paper, we extend the model in [1] to AFS systems and study the relationship between system performance metrics (Vmin/Fmax) and system design parameters such as response delay and PLL parameters. While our proposed model is a straightforward extension of that presented in [1], it allows the exploration of the system-level impact of PLL design parameters such as loop bandwidth and size of the frequency step. Also, we will re-examine the effect of design insights described in [1] with a simple supply droop model to provide analytical flavors to obtained results. Moreover, we aim to highlight the key design considerations regarding bandwidth constraints of each sub-component that can serve as a guideline for AFS system designs.



(a)



(b)

Fig. 2: (a) Simplified PDN Model, (b) PDN Impedance Profile with Resonances Highlighted [11].

II. Background

Popular approaches to mitigate the effect of supply noise can be classified as: Power Delivery Network (PDN) design, adaptive clocking circuits and architectural-level error recovery schemes. The most straightforward one is to place enough capacitors at each board/package/die-level to make supply impedance as low as possible across a wide range of frequencies [12]. However, it is generally impractical to use enough bulky capacitors to completely flatten out supply impedance. Another approach to address the supply noise issue, called Resilient Design is to detect and correct the errors at the architectural state of the processor through error recovery processes such as instruction replay [13]. Though a resilient design approach can resolve the inherent bandwidth limitations of adaptive clocking circuits, it is challenging to cover all the possible scenarios of architectural state corruption. Here we focus on the Adaptive Clocking designs where supply droop mitigation is implemented only inside the clock generation/distribution circuits.

A. Supply Droop

Dominant factor of supply droop is resonance peaks at the power supply impedance, where board/package/die-level inductances and capacitances induce resonance peaks at multiple frequencies and amplify injected noise current from surrounding digital systems as seen from Fig. 2a. Fig. 2b shows the resonance peaks from typical power supply impedance profiles that occur at three distinct frequencies. For example, first-order resonance happens at approximately 100MHz, while second- and third-order resonances at ∼1MHz and ∼10kHz, each associated with $L_{pkg}C_{die}$, $L_{pcb}C_{pkg}$, $L_{reg}C_{bulk}$ respectively [11]. As $\Delta V = \Delta I * Z$, both current noise injection and supply impedance profile determine the magnitude and frequency of supply voltage droop. Hence supply voltage droop is actually a function of workload variations which makes only post-silicon measurements valid for the droop characterizations [12]. Nevertheless, supply droop is often simplified as a combination of the first, second and third droop each corresponding to resonant peak.

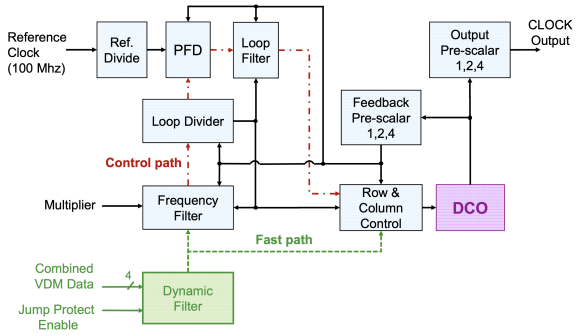
Mitigating supply droop is challenging in both magnitude and frequency perspective. In frequency perspective, the first droop in 100MHz has also the largest magnitude of three and forces the circuits to detect and respond to it in a few nanoseconds. On the other hand, in magnitude perspective, technology scaling drives di/dt to a larger value and hence the larger droop, since di is proportional to the transistor counts and $1/dt$ is proportional to the clock frequencies [14]. Moreover, as described in [15], core-to-core interactions can cause perfect storms of supply droop, which is approximately 1.2x larger in magnitude and 1.9x steeper in slope.

B. State-of-the-art Adaptive Clocking Circuits

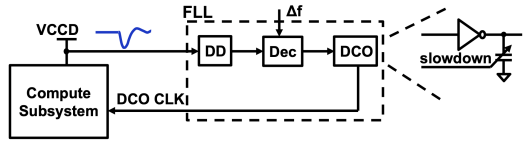
AMD's Zen Processor adopted the CS scheme to change the clock frequency by continuously switching clock phases and stretch the clock at a droop event [2]–[4]. However, the main drawback of this scheme is synchronization overhead to the response latency, where a synchronizer block should be added in between the droop detector and the phase picker logic to safely use the output signal from the droop detector. Also, additional phase picking logic adds jitter to the output clock by 0.5∼1% UI. In Zen Processor, the scheme is revised to have coarse- and fine-grain clock-stretchers to reduce the overall response time.

Another interesting technique proposed by Bowman [5], [6] is ACD, where CDC is exploited to prevent the timing failure for multiple cycles after the onset of supply droop. CDC is a phenomena that the clock edge is pushed out by slow-down of clock distribution path proportionally to slow-down of the datapath delay both at a Vdd droop. With CDC, we can secure a sufficient amount of time to respond to the timing margin degradation which is delayed by several cycles. This scheme can also be auto-calibrated [6], removing tester calibration overheads. Moreover, It is fast as it does not require a synchronizer block between detection and adaptation blocks. However, clock buffers and clock gating cells should be fine-grain controlled which is expensive in terms of clock distribution designs.

Further reduction of response time is possible by implementing the system with proactive response rather than reactive. Recent IBM's z15 chip [8] combined proactive voltage droop detection with traditional Critical Path Monitor to further reduce the detection response time. However, as pointed out in [7], proactive detection is prone to mispredictions due to an unpredictable dataflow execution in general-purpose CPUs. Thus the Qualcomm's PCG system in its recent Hexagon DSP is implemented based on the



(a)



(b)

Fig. 3: (a) AFS by Modulating DCO and Divider by VDM [16], (b) AFS by Modulating DCO Capacitors by SR-DD [17].

assumption that the vector coprocessor with highly predictable dataflow execution dominates the power consumption of the system.

Clock generators like PLL or FLL can be asynchronously modulated to accommodate the instantaneous clock frequency adaptation without incurring much overhead to the design. Also known as Adaptive Frequency Synthesis, the idea was first proposed with analog PLL in Nehalem chip [14], where VCO supply is generated by mixing the regulated analog supply and varying digital supply. As presented in Fig. 3a, IBM also adopted this approach in POWER7 and POWER9 family chip's DPLL design [16], [18], [19] by adding direct modulation path from Voltage Droop Monitor (VDM) output to DCO and Divider input with only 15% area overhead and 6ns of total response latency. Likewise in Fig. 3b, with the benefit of minimal overhead, Intel recently implemented this feature into their DVFS-compliant FLL with Self-Referenced Droop Detector (SR-DD) and full asynchronous path to DCO capacitors, resulting in a total latency of 500ps [17]. In both IBM and Intel's designs, gradual-exit mode from frequency adaptation is also implemented to prevent unintentional overshoot when recovering back to the original frequency. While it has desirable features such as high adaptation bandwidth and small area overhead, its modulation gain should be calibrated across PVT and DVFS modes to make sure the high adaptation bandwidth and PLL stability is always guaranteed.

C. Response Delay Model for Adaptive Clocking

As the response delay of adaptive clocking circuits in modern processors should be faster than a few nanoseconds to adapt the clock frequency to supply droops, response latency should be properly modeled to understand how to

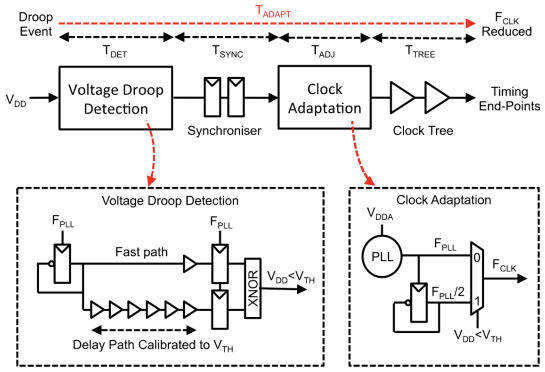


Fig. 4: Response Delay Model in [1].

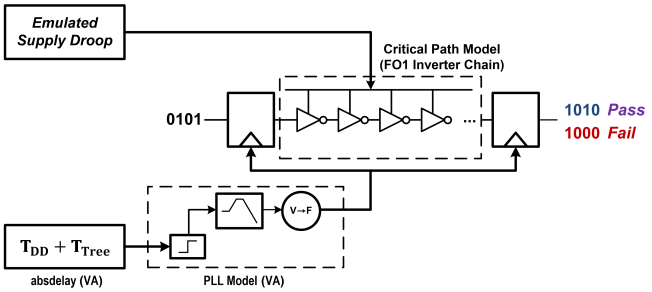


Fig. 5: Proposed Model.

design subcircuits and how to connect them. Assuming a representative system consisting of a droop detector, synchronizer, clock adaptation block and clock tree, [1] models the response delay as a sum of delays at each subcomponent. [1] then identifies the synchronization and clock tree insertion delay as a key limitation to driving the circuits in nanosecond-scale speed. Finally, it is verified that the smaller delay translates into Vmin improvement with the delay model of adaptive clocking using dataset from a specific workload in dual-core ARM Cortex-A57 processor.

III. Proposed Model

Fig. 5 is our proposed model that extend the response delay model in [1] to include PLL dynamics and first-order supply droop. As delay model in [1] is assuming that the droop detection and the clock adaptation is done sequentially and connected via synchronizer, it cannot be used to analyze AFS. To extend the model to the target system, we plan to explicitly include the simplified PLL model into the model and analyze the effect of PLL dynamics. Also, we plan to simplify supply droop behavior as the first-order supply droop only. While the realistic supply droop should include all three orders of supply droop, or should be extracted from the processor model and the actual workload, a simplified droop model will allow us to gain more useful insights into the system. Finally, as the final performance metric is Vmin, we will also use simple models for critical path to emulate timing failures at low Vdd to obtain the minimum achievable supply voltage.

Thus we plan to implement SPICE-based simulation model for AFS that includes:

1) Simplified PLL model implemented in Verilog-A using variable-domain translator block [20]
2) Droop detector response delay T_{DD} and Clock tree insertion delay T_{tree} in Verilog-A absdelay
3) First-order supply droop model implemented in either simplified PDN model or Verilog-A block
4) FO-1 inverter chain to emulate the critical path

IV. Expected Results

With the completion of simulation model, we will attempt to answer the following:

- How does the PLL's loop bandwidth affect Vmin of the system, and how does it related to PSN frequency?
- How does the variations in size of the frequency step used for clock adaptation affect Vmin?
- How critical is the clock tree insertion delay in AFS, which takes approximately 1ns [16]?
- Should droop detector also modulate the clock divider as in [16], [18]?

V. Conclusion

In this paper, we proposed a SPICE-based system-level simulation model for AFS. As the response latency requirements of such systems are becoming more stringent with technology scaling, it is important to design the system that meets such requirements. AFS is one of the techniques that can mitigate the fast supply droop with small response latency and minimal area overhead. To this end, we reviewed state-of-the-art adaptive clocking techniques and modeling efforts that served as a basis of identifying common design practices of the system. Moreover, our proposed model is based on the ARM's response delay model which also played a key role in AMD's clock-stretcher designs. As the model implementation will be detailed in the final writeup, we expect to reveal important design tradeoffs through our system model. A future research direction can include a realistic supply droop model with workload variations, explorations of droop detector and DPLL topologies that can lead to the actual implementations of AFS systems.

References

[1] P. N. Whatmough, S. Das, and D. M. Bull, "Analysis of adaptive clocking technique for resonant supply voltage noise mitigation," in *2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 2015, pp. 128–133.

[2] A. Grenat, S. Pant, R. Rachala, and S. Naffziger, "5.6 adaptive clocking system for improved power efficiency in a 28nm x86-64 microprocessor," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. IEEE, 2014, pp. 106–107.

[3] K. Wilcox, R. Cole, H. R. Fair III, K. Gillespie, A. Grenat, C. Henrion, R. Jotwani, S. Kosonocky, B. Munger, S. Naffziger *et al.*, "Steamroller module and adaptive clocking system in 28 nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 24–34, 2014.

[4] T. Singh, A. Schaefer, S. Rangarajan, D. John, C. Henrion, R. Schreiber, M. Rodriguez, S. Kosonocky, S. Naffziger, and A. Novak, "Zen: An energy-efficient high-performance $\times$86 core," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 102–114, 2017.

[5] K. A. Bowman, C. Tokunaga, T. Karnik, V. K. De, and J. W. Tschanz, "A 22 nm all-digital dynamically adaptive clock distribution for supply voltage droop tolerance," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 4, pp. 907–916, 2013.

[6] K. A. Bowman, S. Raina, J. T. Bridges, D. J. Yingling, H. H. Nguyen, B. R. Appel, Y. N. Kolla, J. Jeong, F. I. Atallah, and D. W. Hansquine, "A 16 nm all-digital auto-calibrating adaptive clock distribution for supply voltage droop tolerance across a wide operating range," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 8–17, 2015.

[7] V. K. Kalyanam, E. Mahurin, K. A. Bowman, and J. A. Abraham, "A proactive system for voltage-droop mitigation in a 7-nm hexagon™ processor," *IEEE Journal of Solid-State Circuits*, 2020.

[8] D. Wolpert, C. Berry, B. Bell, A. Jatkowski, J. Surprise, J. Isakson, O. Geva, B. Deskin, M. Cichanowski, D. Hamid *et al.*, "Cores, cache, content, and characterization: Ibm's second generation 14-nm product, z15," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 98–111, 2020.

[9] K. L. Wong, T. Rahal-Arabi, M. Ma, and G. Taylor, "Enhancing microprocessor immunity to power supply noise with clock-data compensation," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 4, pp. 749–758, 2006.

[10] D. Jiao, J. Gu, and C. H. Kim, "Circuit design and modeling techniques for enhancing the clock-data compensation effect under resonant supply noise," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 10, pp. 2130–2141, 2010.

[11] S. Das, P. Whatmough, and D. Bull, "Modeling and characterization of the system-level power delivery network for a dual-core arm cortex-a57 cluster in 28nm cmos," in *2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 2015, pp. 146–151.

[12] R. Bertran, A. Buyuktosunoglu, P. Bose, T. J. Slegel, G. Salem, S. Carey, R. F. Rizzolo, and T. Strach, "Voltage noise in multi-core processors: Empirical characterization and optimization opportunities," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 2014, pp. 368–380.

[13] K. A. Bowman, "Adaptive and resilient circuits: A tutorial on improving processor performance, energy efficiency, and yield via dynamic variation," *IEEE Solid-State Circuits Magazine*, vol. 10, no. 3, pp. 16–25, 2018.

[14] N. Kurd, P. Mosalikanti, M. Neidengard, J. Douglas, and R. Kumar, "Next generation intel core™ micro-architecture (nehalem) clocking," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1121–1129, 2009.

[15] I. Pierce, J. Chuang, C. Vezyrtzis, D. Pathak, R. Rizzolo, T. Webel, T. Strach, O. Torreiter, P. Lobo, A. Buyuktosunoglu *et al.*, "26.2 power supply noise in a 22nm z13™ microprocessor," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2017, pp. 438–439.

[16] M. S. Floyd, P. J. Restle, M. A. Sperling, P. Owczarczyk, E. J. Fluhr, J. Friedrich, P. Muench, T. Diemoz, P. Chuang, and C. Vezyrtzis, "26.5 adaptive clocking in the power9™ processor for voltage droop protection," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2017, pp. 444–445.

[17] P. Mosalikanti, Q. Wang, K.-Y. J. Shen, M. Neidengard, S. F. S. Farooq, V. Grossnickle, and N. Kurd, "29.3 80ns fast-lock 0.4-to-6.5 ghz clock generator with self-referenced asynchronous adaptive droop mitigation," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64. IEEE, 2021, pp. 408–410.

[18] C. R. Lefurgy, A. J. Drake, M. S. Floyd, M. S. Allen-Ware, B. Brock, J. A. Tierno, and J. B. Carter, "Active management of timing guardband to save energy in power7," in *2011 44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2011, pp. 1–11.

[19] C. Gonzalez, M. Floyd, E. Fluhr, P. Restle, D. Dreps, M. Sperling, R. Rao, D. Hogenmiller, C. Vezyrtis, P. Chuang *et al.*, "The 24-core power9 processor with adaptive clocking, 25-gb/s accelerator links, and 16-gb/s pcie gen4," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 91–101, 2017.

[20] J. Kim, K. D. Jones, and M. A. Horowitz, "Variable domain transformation for linear pac analysis of mixed-signal systems," in *2007 IEEE/ACM International Conference on Computer-Aided Design*. IEEE, 2007, pp. 887–894.