Shortstop: An On-Chip Fast Supply Boosting Technique

Nathaniel Pinckney, Matthew Fojtik, Bharan Giridhar, Dennis Sylvester, and David Blaauw
University of Michigan, Ann Arbor, MI, npfet@umich.edu

Abstract

Fast boosting of supply rails is critical for near-threshold computing to overcome serial code bottlenecks. A novel supply boosting technique, called Shortstop, boosts a 3nF core in 26ns while maintaining acceptable supply voltage droops. The innate parasitic inductance of a dedicated dirty supply rail is used as a boost-converter and combined with an on-chip boost capacitor. Shortstop boosts a core up to 1.8× faster than a header-based approach, while reducing supply droop by 2-7×.

Introduction

Transistor threshold voltages have stagnated in recent technology nodes, deviating from constant-voltage scaling theory and directly limiting supply voltage scaling. To overcome the resulting energy and power dissipation barriers, energy efficiency is improved through aggressive voltage scaling, and recently there is increased interest in operating at "near-threshold" supply voltages [1]. In this region sizable energy gains are achieved with moderate performance loss for parallel applications.

Even for applications that parallelize fairly well, serial portions of code remain. In a near-threshold scenario where most cores run at low voltage, it is therefore advantageous to rapidly increase core voltage to address the need for fast execution of these serial fragments [2] and respond to varying workloads. Such dynamic voltage and frequency scaling (DVFS) is traditionally implemented with an off-chip regulator, which requires hundreds or thousands of CPU cycles to transition to, and stabilize at, a new voltage [3]. Thus, it is not suitable for fast voltage control that responds to fine grain code sequences. To improve performance, on-chip low-dropout regulators have been proposed at the expense of degraded energy efficiency and overall power dissipation. Recent work [4,5] used on-chip regulators to improve speed. Alternatively, DVFS can be implemented on-chip with multiple power rails connected dynamically to a core with PMOS headers. This is faster than off-chip regulators and more efficient than on-chip regulators but incurs voltage droop during transitions, causing timing failures in other cores sharing the power rail.

Technique

We propose a new circuit technique *Shortstop* that addresses power supply droop seen by other cores while boosting a core from a low (0.4V) to high voltage (1.0V) within 26 to 142 ns for ARM M3 or Intel Atom-sized cores, respectively. Shortstop adds a second "dirty" supply rail and an on-chip boost capacitor to rapidly boost the core. The key idea is to transition the cores to high voltage using the dirty supply, thereby decoupling the transition from the clean high voltage supply and isolating other cores from supply droop. In addition, we use the dirty supply's wirebond/C4 innate parasitic inductance in a boost converter arrangement, thereby exploiting this inductance as an asset rather than barrier to fast supply transitions. Finally, on-chip decoupling capacitance is configured as a boost capacitor, further aiding supply transition. The boost capacitor and additional dirty supply are shared between multiple processors, amortizing their overhead.

The key challenge in Shortstop is to boost the supply quickly without destabilizing the power rails used by other cores. A PMOS header implementation has three drawbacks: 1) unavoidable wirebond/C4 inductance creates droop and ringing during fast switching; 2) droop must be small (e.g., <10%) so cores sharing a power rail are not disturbed; 3) adding on-chip decoupling capacitance to the power rail to reduce droop and ringing incurs large area costs.

Shortstop addresses these issues through the use of dirty VDD, *Vdirty*, (Fig. 1). *Vdirty* is connected to the high supply voltage (e.g., 1V) off-chip and does not require additional off-chip regulation. Since the *Vdirty* supply is used only for transitioning a single core at a time, it does not need to be as robust as nominal operating supplies. This greatly reduces overhead since *Vdirty* need only use a small number of pads that are amortized across many cores. Shortstop consists of one header block per core and a single shared boost block. Fig. 2 describes the basic operation of Shortstop. A core is initially connected to the low voltage supply *Vlow* and the on-chip capacitor is charged to *Vhigh* (Step 1). The core is then switched from *Vlow* to the capacitor, which partially boosts the core while *Vdirty* is simultaneously shorted to a dirty ground to energize *Vdirty*'s inductance, similar to boost converter operation (Step 2). The on-chip capacitor must be large, similar to the intrinsic capacitance of a core, but is shared across several cores to reduce area overhead. Once charge sharing is complete the core is switched from the capacitor to *Vdirty* supply (Step 3). Since by this time, significant energy has built up in the *Vdirty* inductor, *Vdirty* quickly boosts the core to full voltage.

As the core reaches the target high-voltage supply, it is switched from *Vdirty* to the nominal *Vhigh* supply (Step 4). Since the core is already charged to a level near *Vhigh*, this step does not incur significant droop or otherwise destabilize the *Vhigh* supply, which thus can be shared by a number of other cores. However, *Vdirty* will incur significant ringing and actually overshoot the high supply voltage, which is undesirable. Two techniques are used to avoid this ringing: 1) when *Vdirty* is disconnected from the core, it is immediately connected to the on-chip capacitor to use the remaining wirebond/C4 inductance energy to charge the capacitor, preparing it to boost another core. When *Vdirty* has transferred its energy to the boost capacitor and reaches its maximum voltage, *Vdirty* is disconnected and clamped to *Vhigh* to immediately suppress any further ringing (Step 5). Since *Vdirty*'s inductor has was discharged when clamped, and Vdirty has no on-chip decoupling capacitance, this step does not disturb *Vhigh*.

The Shortstop boosting steps must be timed accurately (100s of ps) to function efficiently. This is accomplished using programmable on-chip delay generators that are tuned for a particular package and chip configuration. Alternatively, high-speed comparators [6] could be used in an automated timing architecture. The on-chip timing circuitry includes a 1.25 GHz asynchronous clock generated by a ring oscillator, 16 delay generators with fine (25 ps) and coarse delay (800 ps) steps, and maskable XOR trees that combine multiple timing signals into arbitrary digital waveforms for the switches. The test chip architecture includes two on-chip variable capacitors/current sources to emulate large cores, as well as an actual implemented ARM Cortex M3 core. In addition to timing control circuits, headers, and a boost capacitor, on-chip samplers monitor power rails using a sample-and-hold averaging technique [7] that enables an effective bandwidth of ~40 GHz.

Measured Results

Shortstop is validated in a 28nm CMOS test chip measuring 3.9 mm^2 (Fig. 3). The chip is wirebonded to an 88-pin QFN package and a 108-pin ceramic PGA package with two wirebond lengths to vary package parasitics. Fig. 4 shows silicon measurements comparing boosting time for the included M3 core using a baseline PMOS header based approach and Shortstop. The 1-pin baseline assumes Shortstop's hardware overhead can be amortized across multiple cores and hence is negligible, while the 2-pin baseline is a conservative estimate where the number of dirty supply pins equals the number of high supply pins. For the M3 core, boost latency and droop are improved by 1.7× and 6×, respectively.

Fig. 5 compares supply droop and boost latency, defined by rise time within 10% of Vdd, for the baselines and Shortstop across different emulated core sizes. As core size decreases, Shortstop exhibits slightly increased gains against baselines, while supply droop is relatively constant at 6× and 3× for the 1-pin and 2-pin baseline, respectively. For a 15 nF core (an Intel Atom-sized core), boost latency is improved by 1.6× in addition to a 6× droop reduction. Fig. 6 shows the impact of boost capacitance size on supply droop and latency indicating that 30–40% of intrinsic core capacitance is sufficient to obtain most of Shortstop's performance gains. Finally, Fig. 7 shows Shortstop maintains a 1.4× latency improvement and 4× droop reduction across the three packages tested. As package parasitics decrease, the baseline latency and droop improves but this is balanced by decreased parasitics on the dirty supply which shortens boost time to energize the parasitic inductance.

Acknowledgement

STMicroelectronics is gratefully acknowledged for IC fabrication.

References

[1] R. G. Dreslinski et. al., Proceedings of the IEEE, February 2010.

[2] T. N. Miller, HPCA, 2012.

[3] E. Rotem et. al., MICRO, March 2012.

[4] P. Hazucha et. al., IEEE Journal of Solid-State Circuits, April 2005.

[5] W. Kim et. al., ISSCC, February 2011.

[6] S. Pant et. al., ISSCC, 2008.

[7] R. Ho et. al., Symposium on VLSI Circuits, 1998.

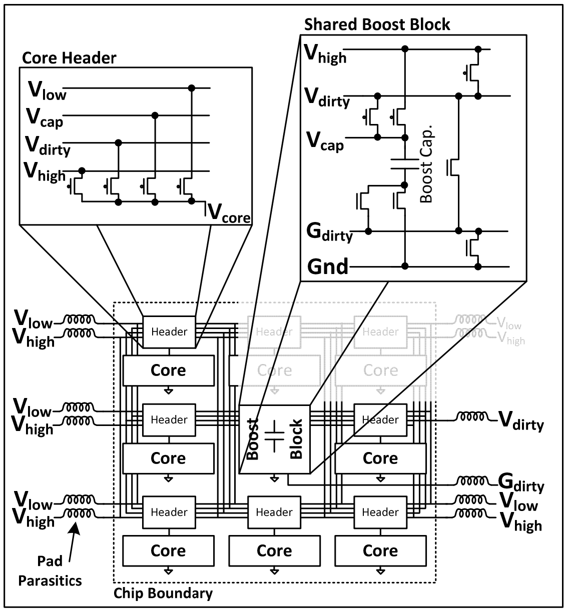
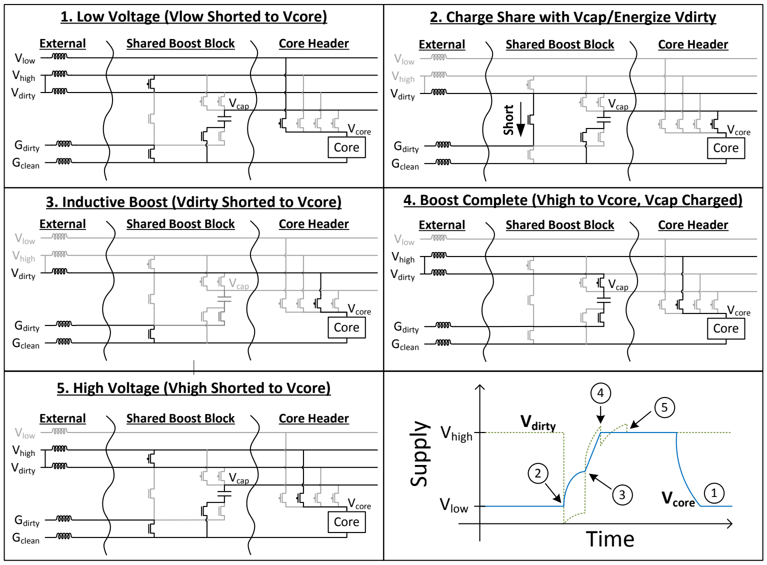
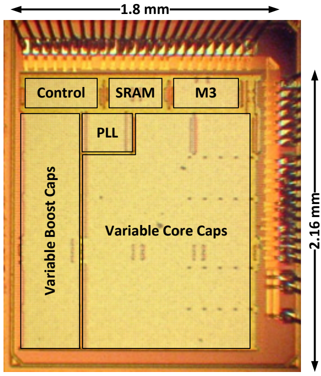


Figure 1. Shortstop architecture.



1. Low Voltage (Vlow Shorted to Vcore)
2. Charge Share with Vcap/Energize Vdirty
3. Inductive Boost (Vdirty Shorted to Vcore)
4. Boost Complete (Vhigh to Vcore, Vcap Charged)
5. High Voltage (Vhigh Shorted to Vcore)
External
Shared Boost Block
Core Header
V_{low}
V_{high}
V_{dirty}
V_{cap}
V_{core}
Core
G_{dirty}
G_{clean}
Short
Supply
V_{high}
V_{dirty}
V_{low}
V_{core}
Time
1
2
3
4
5

Figure 2. Steps of Shortstop to boost core supply rail.



Shortstop Test Chip Specifications	
Technology	28 nm
Area	3.9 mm^2
Processor Core	ARM Cortex-M3
Max Core Cap.	15 nF
Max Boost Cap.	5 nF
Package Variants	88-Pin QFN (short bond wires) 108-Pin CPGA (med. bond wires) 108-Pin CPGA (long bond wires)

Figure 3. Photomicrograph of 28nm test chip and chip specifications.

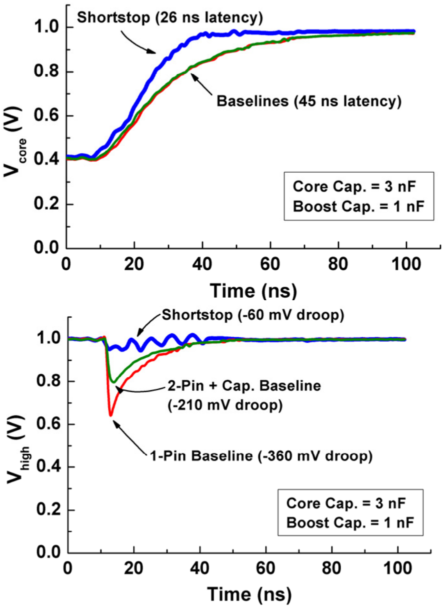


Figure 4. Measured rail voltages for 3 nF core (QFN package).

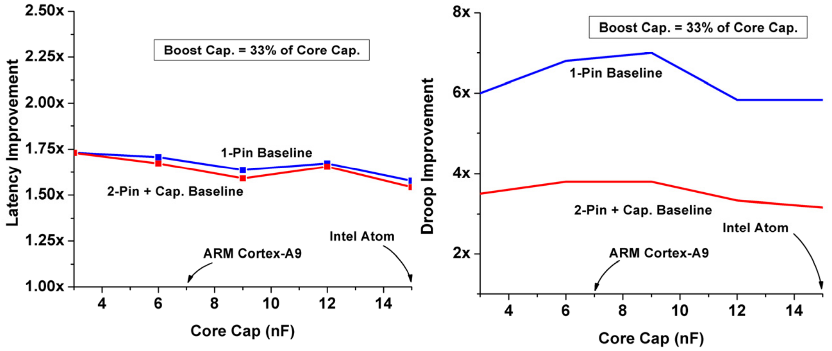


Figure 5. Measured latency/droop improvement for varying core cap. (QFN package).

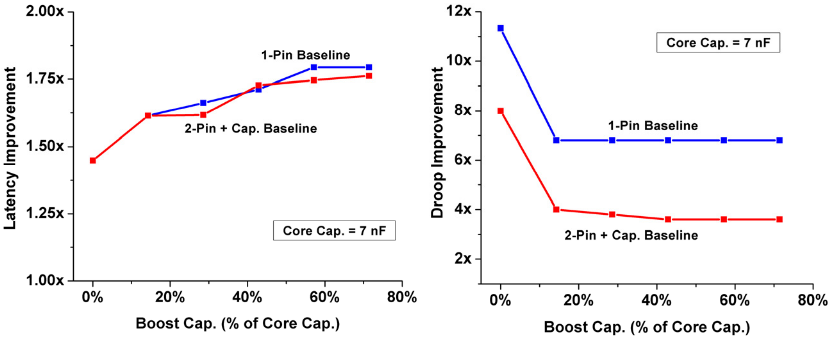


Figure 6. Measured droop and latency for varying boost cap. (QFN package).

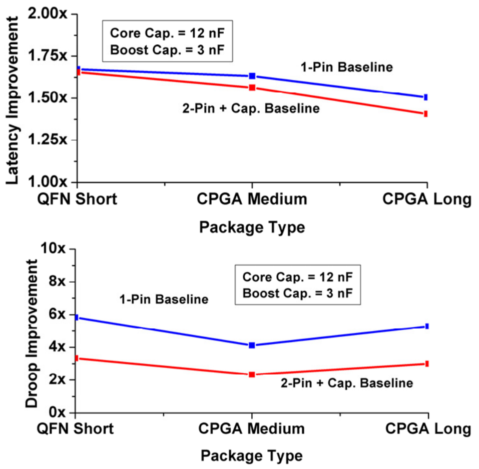


Figure 7. Measured performance improvements with varying packages and wirebond lengths.