Supply Boosting for High-Performance Processors in Flip-Chip Packages

Nathaniel Pinckney, Dennis Sylvester, and David Blaauw
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI
npfet@umich.edu

Abstract—**On-chip supply boosting can quickly restore a microprocessor core's power rail from near-threshold to super-threshold when critical code sections are encountered. We demonstrate a flip-chip implementation of a supply boosting technique, called Shortstop, which uses a transient supply rail and leverages the parasitic and intentional inductance of a package. To address package parasitic variation, an automatic tuning algorithm is shown. A 7.9mm^2, 40nm CMOS prototype chip is attached to a custom ball grid array substrate, with integrated in-package inductors. Shortstop boosts a 2.7mm^2 core from 0.5V to 0.75V in 14ns with only 27mV of droop on a shared 0.8V supply rail, marking a 57% faster transition with 67% lower supply noise than a dual-supply PMOS header design.**

Keywords—Supply boosting, near threshold, low power, dual rail.

I. INTRODUCTION

Today's power scaling limitations have led to increased use of dynamic voltage scaling to near-threshold voltages [1]. To accommodate changing workloads and supply voltage requirements, dynamic boosting schemes [2-4] are useful to quickly adjust the operating voltage of a core and maximize energy efficiency by matching supply voltage to workload. As code transitions from energy-efficient parallelized execution to serial, a core's voltage requires rapid escalation to achieve high single-thread performance.

Shortstop [2] is a previously-proposed technique that uses PMOS power headers, three supply voltage rails (one low voltage, two high voltage), and an on-chip capacitor to raise core voltage. The time to boost, or boost latency, is much faster than in the case of off-chip voltage regulation and unlike on-chip regulators [4] it does not require special in-package or in-die materials to achieve high efficiency. Furthermore, it leverages package parasitic inductance to boost the supply inductively, minimizing boost latency.

In [2] a wirebonded implementation was demonstrated for Shortstop, but modern high-performance microprocessors are packaged in flip-chip technologies. This work expands upon previous Shortstop work by: (1) demonstrating Shortstop in a custom BGA flip-chip package and showing how in-package inductors can reduce boost latency; (2) introducing a new on-chip self-tuning algorithm that address package parasitic variations; and (3) a new architecture and physical design strategy.

II. PROPOSED DESIGN

Fig. 1 shows the proposed Shortstop architecture, which removes one power header from the core at the cost of two power headers in the shared boost block (circled) compared to prior work. In [2] the on-chip boost capacitor rail V_{cap} and transient supply rail V_{dirty} were distributed over the cores, while in the proposed design they are multiplexed in the shared boost block and distributed via a V_{boost} virtual supply rail. Since there are many more cores than shared boost blocks, less area is consumed in the core area for power switches.

Fig. 2 shows switch status with each step of a boost. Initially the core's virtual supply rail V_{core} is connected through a power switch to the low supply rail V_{low} (0.5V in the example). When a boost request is received, the system connects V_{core} to the transient boost supply rail V_{boost}, which is connected to V_{cap} in the shared boost block and pre-charged to 0.8V. Connecting V_{cap} provides an initial pre-charge to the core while simultaneously the external transient supply V_{dirty} is shorted to ground in order to energize its associated package inductance. In Step 3, V_{cap} is disconnected from V_{boost} and connected to V_{dirty} instead, rapidly boosting V_{core} with energy

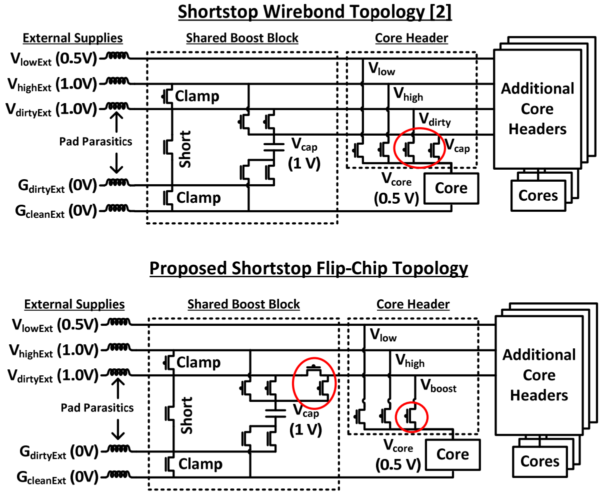


Fig. 1 Proposed Shortstop topology compared to wirebond Shortstop topology from [2]. Modifications to Shortstop architecture are circled.

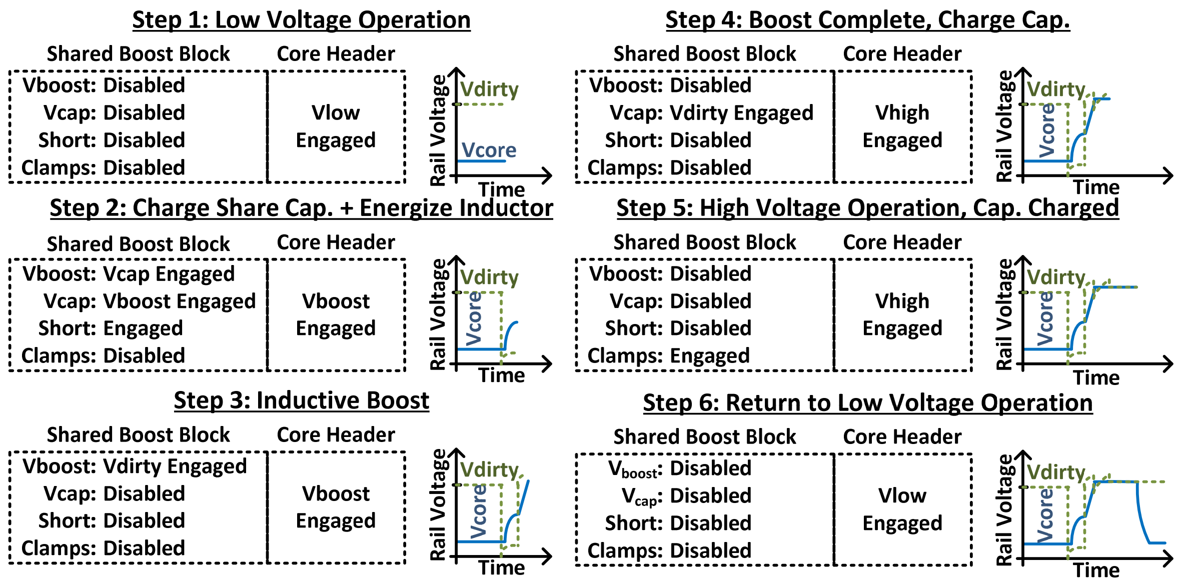


Fig. 2 Steps of Shortstop FC technique.

from the inductor until the target voltage (near 0.8V) is reached, at which point V_{boost} is disconnected and the primary high supply rail V_{high} is connected. As V_{core} nears V_{high}, very little ringing is introduced when this final connection is made in Step 4. The remaining Steps 5 and 6 reset the system for another boost.

The previous Shortstop design relied on hand tuning of delay generators to time the power switches for boosting. We propose a digital, automatic tuning approach that uses a clocked comparator in the core area and a finite state machine in the test harness. Fig. 3 details the automatic tuning algorithm, which first measures the time when the V_{high} rail is above a target threshold (set using an off-chip voltage reference). Then using gradient descent it adjusts the boost short and share times until the time to the target voltage is minimized. The comparator voltage reference is set ~30mV below V_{high} to minimize V_{high} droop at the end of the boost cycle.

The core area was split up between sixteen distinct power domains of varying sizes, shown in Fig. 4, allowing different boosting scenarios to be tested. Each core area can be connected to V_{low}, V_{high}, or the supply undergoing a boost between the two rails. Top-layer metal power stripes were designed to minimize the impact of connecting to flip-chip bumps. The 200µm-pitch bump pattern includes V_{low}, V_{high} and V_{SS} bumps over core areas, while the V_{dirty}/G_{dirty} transient supplies and shared boost block are confined to the center of the chip and shared among cores. The test harness and shorting blocks were oversized to align with bump boundaries. Power switches were distributed across core area on a chip, not unlike standard power gated designs. This comes with the disadvantage of requiring clock tree synthesis (CTS) to distribute power switch enable signals with minimal skew to prevent short circuit currents between power rails.

Each core area includes a test island with samplers for observability and a clocked comparator for digital tuning. Within the core are CMOS filler cap and analog-controlled NMOS current sources between the virtual core supply rails and ground, which serve to emulate the capacitance and power draw of a core.

III. Measured Results

The proposed Shortstop architecture was implemented in TSMC 40nm CMOS. Fig. 5 shows a die shot, photo of the test chip attached to the BGA package in a test socket, and test chip summary table. A custom flip-chip BGA package was used to connect the flip-chip die to a PCB for testing, through a BGA socket. The custom package includes four V_{dirty} supply rail connections, one with a straight metal trace and three looped metal traces to add inductance of various sizes (0.5nH, 1nH, and 2nH). Inductance was extracted using Ansys HFSS modeling of the package substrate dielectric and copper traces.

Fig. 6 shows measured on-chip waveforms of V_{core} and V_{high} supply rails for Shortstop and baselines. The baseline configurations connect V_{core} directly to V_{high} without first connecting to V_{boost}. In one baseline we current starve the PMOS V_{high} header. A physical implementation limitation prevented the on-chip boost capacitor from being disconnected from V_{boost}, however the ground side was disconnected through footers with some parasitic capacitance remaining. To compensate, Shortstop results are penalized by turning off the boost capacitor footers, while the footers are on and connected V_{boost} to V_{high} in the baseline case to improve its power delivery. Despite this penalty, Shortstop boosts the core from 0.5V to 0.8V (with −50mV allowable droop) in 14.2ns, versus 20.6 − 32.4ns in the baseline. Even with current starving, the baselines exhibit 82−120mV of droop on V_{high}, while Shortstop suffers just 27mV of droop.

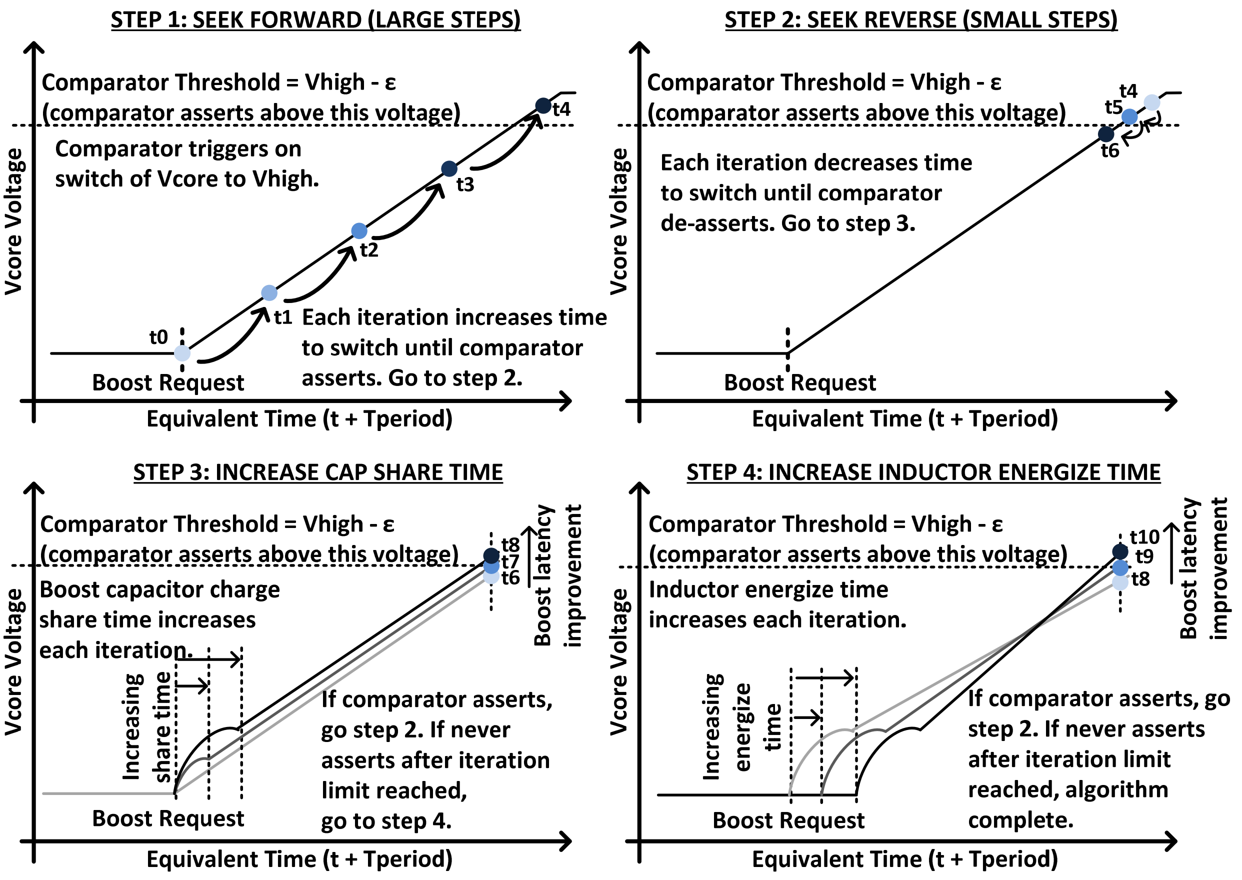

STEP 1: SEEK FORWARD (LARGE STEPS)
Comparator Threshold = Vhigh - ε
(comparator asserts above this voltage)
Comparator triggers on switch of Vcore to Vhigh.
Each iteration increases time to switch until comparator asserts. Go to step 2.
t0
t1
t2
t3
t4
Boost Request
Vcore Voltage
Equivalent Time (t + Tperiod)
STEP 2: SEEK REVERSE (SMALL STEPS)
Comparator Threshold = Vhigh - ε
(comparator asserts above this voltage)
Each iteration decreases time to switch until comparator de-asserts. Go to step 3.
t4
t5
t6
Boost Request
Vcore Voltage
Equivalent Time (t + Tperiod)
STEP 3: INCREASE CAP SHARE TIME
Comparator Threshold = Vhigh - ε
(comparator asserts above this voltage)
Boost capacitor charge share time increases each iteration.
Increasing share time
If comparator asserts, go step 2. If never asserts after iteration limit reached, go to step 4.
t8
t7
t6
Boost latency improvement
Boost Request
Vcore Voltage
Equivalent Time (t + Tperiod)
STEP 4: INCREASE INDUCTOR ENERGIZE TIME
Comparator Threshold = Vhigh - ε
(comparator asserts above this voltage)
Inductor energize time increases each iteration.
Increasing energize time
If comparator asserts, go step 2. If never asserts after iteration limit reached, algorithm complete.
t10
t9
t8
Boost latency improvement
Boost Request
Vcore Voltage
Equivalent Time (t + Tperiod)

Fig. 3 High-level steps of automatic tuning algorithm to minimize boost transition time. The algorithm employs a gradient descent while increasing Shortstop charge share and inductor energization times.

Fig. 7, top, shows measured Shortstop performance versus core size with an in-package inductor of 2nH, when boosting from 0.5V to 0.75V. Compared to the baseline, Shortstop reduces boost time by 36% to 56% for a 2.5mm^2 core while not exceeding 31mV of V_{high} droop (82−120mV of droop for the baseline). The smallest core size tested, 0.64mm^2, has a boost time of 7.8ns, while the largest core of 2.7mm^2 boosts in 14.2ns.

Shortstop partially relies on V_{dirty}'s parasitic inductance to improve boost latency, so sweeps were measured with and without shorting of V_{dirty} prior to connecting to V_{core} through V_{boost}. Fig. 7, bottom, shows performance versus in-package inductor with core voltage areas one through ten activated (equivalent to a core size = 2.7mm^2). Shorting introduces an additional 4mV of droop on V_{high}, but did not exceed 35mV of total droop for the sizes measured. When energizing V_{dirty}'s parasitic inductance, boost time improves by up to an additional 11%, indicating that using the transient rail alone adds a substantial improvement to boost latency when limiting supply noise in flip-chip designs.

In summary, a core supply rail boosting technique, called Shortstop, is demonstrated in a custom flip-chip package. The design improves upon prior work [2] through a new proposed architecture, a distributed and modular physical design approach applicable to flip-chip microprocessors, and an automatic tuning FSM. Shortstop in flip-chip improves upon a dual-rail PMOS header-based technique with 57% faster transition time and 67% lower supply noise.

Acknowledgment

The authors thank Oracle for tapeout support, SRC for the Shortstop concept development, and Advotech for packaging.

References

[1] B. Zhai, R. G. Dreslinski, D. Blaauw, and T. Mudge, "Energy efficient near-threshold chip multi-processing," in *Proc. ISLPED*, 2007.

[2] N. Pinckney, M. Fojtik, B. Giridhar, D. Sylvester, and D. Blaauw, "Shortstop: An on-chip fast supply boosting technique," in *Proc. VLSI Circuits*, 2013.

[3] Z. Toprak-Deniz et al., "Distributed system of digitally controlled microregulators enabling per-core DVFS for the POWER8 microprocessor," in *Proc. ISSCC*, 2014.

[4] E. A. Burton, G. Schrom, F. Paillet, J. Douglas, W. J. Lambert, K. Radhakrishnan, and M. J. Hill, "FIVR—fully integrated voltage regulators on 4th generation Intel Core SoCs," in *Proc. APEC*, 2014.

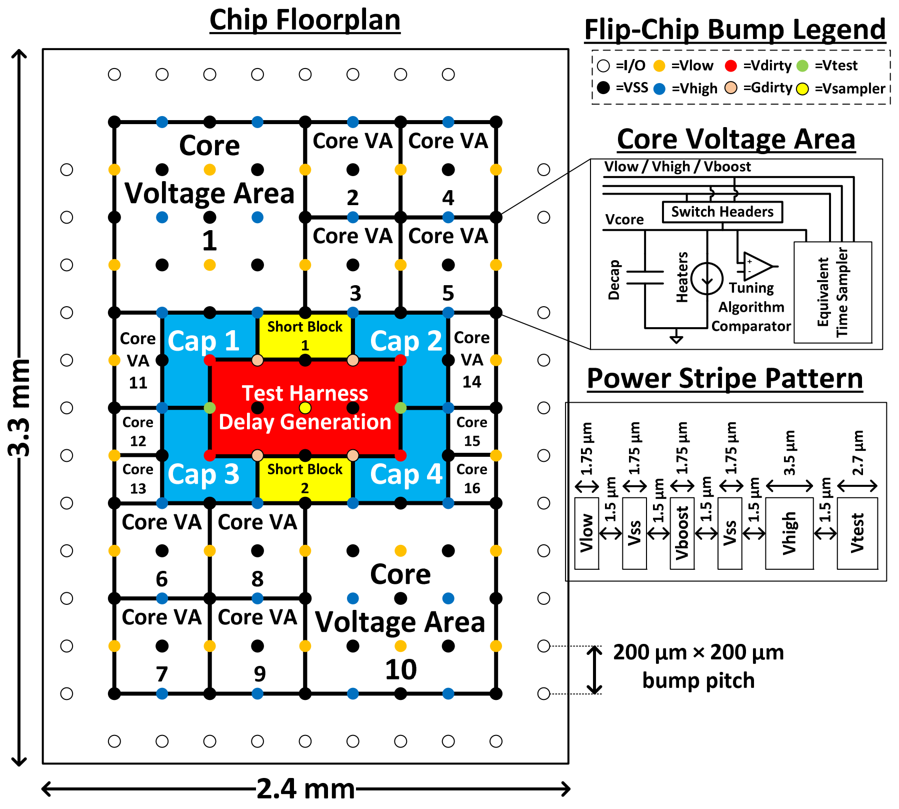
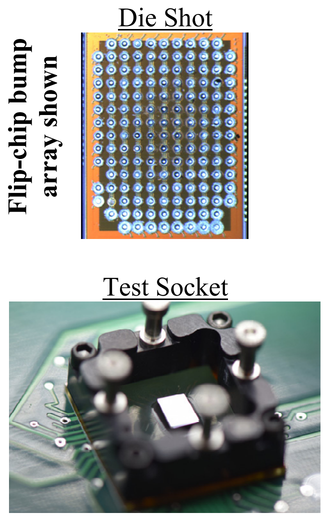


Fig. 4 Shortstop flip-chip modular floorplan concept. Configurable core areas emulate different boosting scenarios and different sized cores.



Chip Specs	
Process	40nm
Area	7.9mm^2
Package	BGA
#Bumps	157
#Balls	72

Fig. 5 Die photo (top), packaged die (middle), and chip specs (bottom).

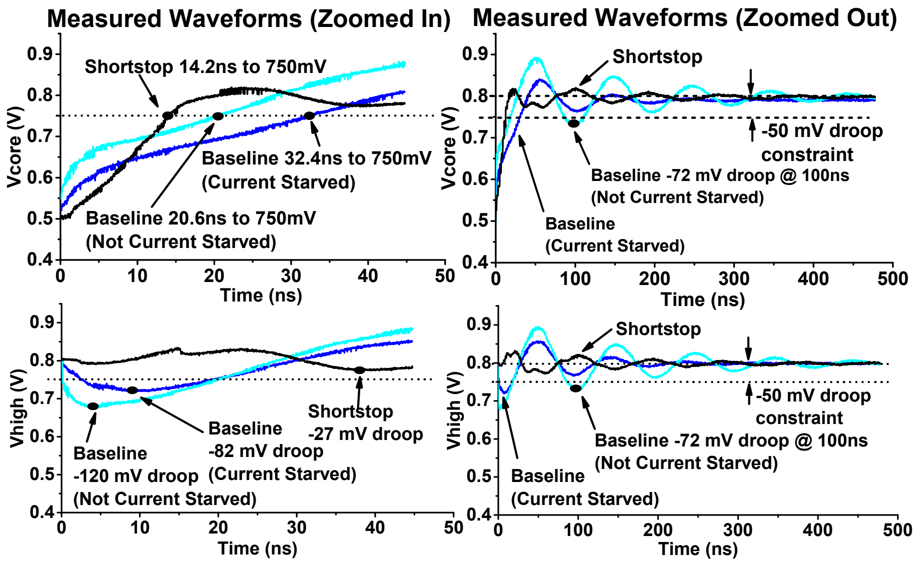


Fig. 6 Measured on-chip waveforms of proposed Shortstop flip-chip architecture versus baseline of a dual-rail PMOS header-based baseline architecture.

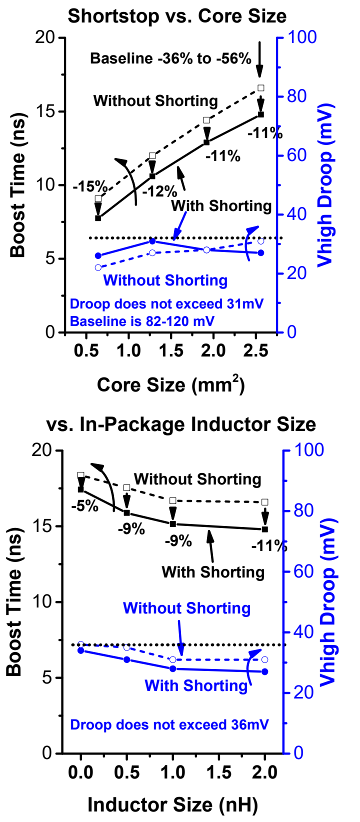


Fig. 7 Shortstop transition time (0.5V to 0.75V) and V_{high} droop versus core size and in-package inductor size.