Multiple Clock and Voltage Domains for Chip Multi Processors

Efraim Rotem
Intel Corporation, Israel
efraim.rotem@Intel.com

Ran Ginosar
Technion, Israel Institute of Technology
ran@ee.technion.ac.il

Avi Mendelson
Microsoft R&D, Israel
avim@microsoft.com

Uri Weiser
Technion, Israel Institute of Technology
uri.weiser@ee.technion.ac.il

ABSTRACT
Power and thermal are major constraints for delivering compute performance in high-end CPU and are expected to be so in the future. CMP is becoming important by delivering more compute performance within the power constraints. Dynamic Voltage and Frequency Scaling (DVFS) has been studied in past work as a mean to increase save power and improving the overall processor's performance while meeting the total power and/or thermal constraints. For such systems, power delivery limitations are becoming a significant practical design consideration, unfortunately this aspect of the design was almost ignored by many research works. This paper explores the various possible topologies to build a high end multi-core CPU and the available policies that maximize performance within the set of physical limitations. It evaluates single and multiple voltage and frequency domains and introduces a new clustered topology, grouping several cores together. A hybrid model, using measurements of a real CPU, cycle accurate simulator and an analytical model is introduced. The results presented indicate that considering power delivery limitations diverts the conclusions when such limitations are ignored. This paper shows that a single power domain topology performs up to 30% better than multiple power domains on light-threaded workload. In the fully threaded application the results divert. Clustered topology performs well for any number of threads.

Categories and Subject Descriptors
C.4 [Performance of Systems]: Design studies

General Terms
Design, Performance, Measurement

Keywords
Power management, DVFS, Clock domains, Voltage domain, Chip Multi Processor

1. INTRODUCTION
Power and thermal limitations force all modern processors to change their design target from frequency driven to multiple cores on die (CMP).Any design of a modern architecture aims to achieve maximum performance within the given power and thermal constraints. On one hand, computer platforms are typically designed for the maximum possible workload, on the other hand, most of the modern CPUs are thermally limited, meaning that could run faster if power delivery and thermal allows. Multi-threaded and multi-process workloads running on CMP often do not stress the CPU to its maximum power. As a result, such systems have power and thermal headroom that could be utilized to extract higher performance out of the system. Dynamic Voltage and Frequency Scaling (DVFS) was proven to be a powerful tool for saving power and thermal, and for achieving higher performance within the power limits. At the same time, demand for single thread and light threaded performance, either for single threaded applications or for serial portions of a multi-threaded workload, remains a challenge. Method that can identify the headroom the parallel execution leaves and translate it into higher performance of the (partially) sequential part can be a key for the performance of many future architectures/workloads.

In this paper we try to maximize total CPU performance within a set of physical constraints. We evaluate several possible clock and voltage domain topologies and management policies. We use the term topology to describe the hardware construction and partition of the clocks and voltage domains of a CPU. The term policy is used to describe the set of algorithms and operating modes used in each given topology. Each study scenario consists of a combination of a topology and a policy. We evaluate different topologies and policies in order to identify the best way to construct the CPU voltage and clock domains, and to manage them for maximizing performance, given a set of physical constraining parameters. In our study, the physical constraints that limit the CPU performance are:

- Maximum voltage and frequency which are a function of the process technology
- Total power consumption and maximum power delivery capability which are platform and package limitations.
- We also account for the minimum voltage and frequency. Minimum voltage is determined by process technology and

minimum operation frequency limitation is driven by clocking design and the minimum required quality of service.

One may assume that adding many power plans (that allow independent voltage and frequency setting for each core) to a die should provide the maximum flexibility. On the other hand, allowing that may incur certain costs. Separating the clock and voltage domains for each core requires clock synchronizers that add latency to the memory and cache transfers, results in degraded power delivery network and is more limited in power frequency scaling. These implications eventually translate into performance degradation, and are not equal for all topologies.

This paper makes three primary contributions. First, we introduce power delivery as a significant constraint to a high power CMP. This constraint is often ignored in multi-core DVFS studies. Accounting for power delivery physical properties results in different conclusions than presented in some prior CMP studies. We show that power delivery limits the benefits of per-core DVFS and that the limitation is especially noticeable in partially threaded workloads. Second, we perform a comprehensive study of different possible topologies of a CMP and study the policies that best utilize these topologies. We evaluate wide range of parameters and show, in contrast to results of previous studies, that independent voltage and frequency domains do not always provide the best performance. We also introduce a clustered topology that has not yet been studied in the context of power management. We show the benefits of such a topology and the considerations of optimal cluster size. Third, we develop a statistical methodology and Monte-Carlo simulation tools to evaluate CMP with a large number of cores. A common practice in evaluating and presenting CMP workload is to use a small number of applications or benchmarks. On CMP with a large number of cores there is large number of possible workload combinations and therefore this approach is very limiting. We show that different workloads can yield different conclusions and considering average results provides only a partial picture.

The rest of the paper is organized as follows: Section 2 survey previous work, section 3 describes our methodology and the results are elaborated in Section 4 and concluded in Section 5.

2. Related Work

Prior research investigated managing CMP power and frequency to maximize performance or energy metrics under a set of physical constraints. Isci et al. [1] studied power management policies on an IBM POWER 4/5 for maximizing performance while observing a power budget. They concluded that a DVFS policy for the individual cores performs significantly better than chip wide DVFS. Power delivery was not considered as a constraint. This paper shows that implementing DVFS per core and accounting for power delivery lead to different conclusions. Ogras et al. [2] employed multiple voltage and clock domains and evaluated the latency and energy cost of clock synchronizers. They targeted energy savings when the chip operates below the nominal operating point. Power delivery was not a constraint and the energy implication of separating power delivery was not accounted for.

Several works collected real time characteristics of the workload. Bellosa at el. [3] collected architectural events and predicted CPU power at run time, using power to perform energy-aware scheduling. Contreras & Martonosi [4] demonstrated a similar concept on Intel® Xscale® Processor. Both studies showed distinct phases during which the CPU remained stable before transitioning to a different phase. Choi et el. [5] implemented a power management algorithm on Intel® Xscale® by observing memory-bound phases of a workload. A hardware mechanism tracked CPU behavior; when a CPU accesses external memory there is little performance gain in running at high frequency because the execution is memory-bound. The algorithm reduced voltage and frequency in such phases to save energy. Asu & Feng [6] described a Performance Monitoring Unit to track memory-bound phases and Wu at el. [7] employed a dynamic compiler to detect these phases in a workload. This paper characterizes the power and scalability of workloads offline, and uses these features to affect power management.

Several studies of power management of CMPs investigated multi-threaded workloads. Isci et al. [1] recognized the importance of single threaded workloads but still evaluated multi-threaded workloads that fully utilized all cores. Hill and Marty [8] studied the implications of Amdahl's law on parallel workloads on both symmetric and asymmetric CPU architectures. Annavaram et al. [9] and Grochowski et al. [10] studied asymmetric CMP under power constraints as a means of mitigating Amdahl's law in CMP. Applying DVFS to CMP allowed the asymmetric operation of a symmetric architecture, e.g. by assigning higher power budget to one of the cores in order to achieve higher single thread performance. These studies indicated that CMPs need to address both single and multi-threaded workloads, as done in this paper.

Joseph et al. [11] considered the architectural impact on the power delivery network and studied control mechanisms that mitigate the impact of voltage variations due to current variations (dI/dt) caused by the activation of power management techniques. Gupta et al. [12] investigated the effect of dI/dt on high frequency components of the power delivery network in CMP. They proposed a modeling infrastructure and evaluated a single voltage domain CMP. These papers, however, considered neither multiple power domains, nor the implications of power delivery on the efficiency of power management algorithms. Kim et al. [13] studied novel integrated on die voltage regulators capable of supporting multiple power domains. The paper focused on the power and energy benefits of fast DVFS response time, but did not study the implications of current delivery constraints. None of these works considered either maximum current or power delivery implications on CMP topology and power management policy; the present paper addresses these issues.

3. Experimental Methodology

This section describes how the reported measurements were conducted, and the next section explains the results.

3.1 Micro architectural model and topologies

The model studied in this paper comprises 16 cores CMP. Each core is identical to the processor used in the Intel® Core™ 2 Duo [14] [23]. The cores are connected via a network-on-chip to a shared L2 cache and to an off-chip memory. The micro-architecture is presented in Figure 1. The following alternative topologies are considered:

1. A single clock domain topology operates all the cores, the interconnect and the shared cache at the same clock frequency. No FIFO buffers are required to synchronize any clock domain crossings. Single clock domain implies also single voltage domain.
2. Multiple clock domains with multiple voltage domains topology that operates each core at different independent frequency, allowing DVFS of each individual core. The interconnect operates at the highest available frequency and voltage. Synchronizing independent clock domains requires FIFO buffers that introduce additional latency into cache and memory data transfers. The latency is a function of the individual frequencies. For simplicity of modeling, an average latency of three clock cycles is added in each direction. .
3. Another possible topology is a multiple clock domain topology with single voltage domain allowing DVFS to all cores simultaneously and DFS to individual core.

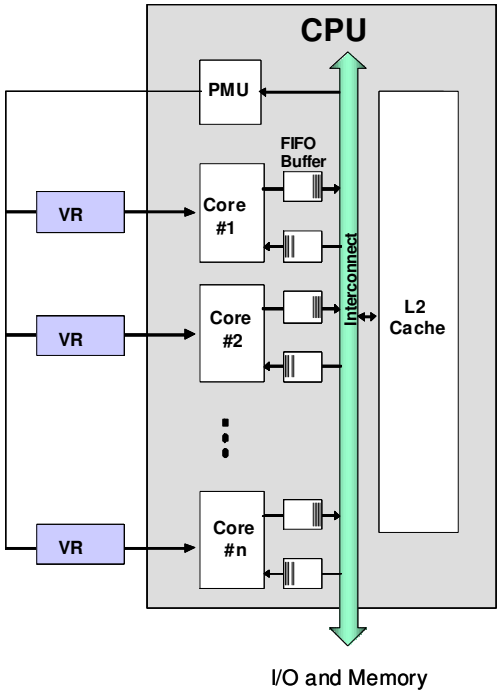

Figure 1: CMP micro-architecture

A multiple voltage domains multiple clock domains allows scaling both voltage and frequency and results in cubic power scaling (P~FV2~V3). All cores are symmetric and on die process variations are ignored [15] and hence different symmetric cores running at the same frequency are supplied the same voltage. Inactive cores are disabled and power gated, turning off leakage. Such power gating is implemented in the Intel® i7 [16] The Power Management Unit (PMU) collects real time information about the executed workload and determine the best policy for maximizing CMP performance, setting the frequency and voltage for each frequency and voltage domain. The PMU communicates with and external voltage regulator(s) (VR) to set the operating voltage(s).

3.2 Power Delivery

Two power delivery schemes are included in this study: a single VR that connects to a single voltage domain, or 16 separate VRs that are connected to the 16 individual cores and an additional separate VR that controls supply to the interconnect and the shared cache. A switching VR comprises a controller, CMOS drivers, inductors and capacitors [17] The passive components need to be placed close to the CMP in order to reduce parasitic resistance and inductance [20] in the power delivery networks (PDN).

Power delivery imposes several constraints. The amount of current delivered by each VR is limited. A VR is designed for a certain nominal current; higher current requires additional or bigger components, increasing cost and incurring more board space close to the CMP, a requirement that may be hard or impossible to meet.

The advantage of a single voltage domain is the capability of sharing the current among the cores; when some cores consume less current or are turned off, current can be directed to the other active cores. This advantage comes at the cost of tying all the voltage domains together, forcing the same operation voltage to all cores. On the other hand, multiple voltage domains topology provides the ability to deliver individual voltages and frequencies according to an optimization algorithm. In particular, when a single thread workload is executed, the entire CMP power budget can be assigned to a single core, which can consume 16 times higher power than each individual cores when executing a balanced workload on all 16 cores. While in both cases the total CMP power is the same, separate power domains require at least one of the 16 VRs to deliver 16 times higher power than its nominal working point. Such a requirement is not feasible. The present study evaluates VR designed to deliver 130%–250% of the rated CMP current.

Another parameter affecting single vs. multiple VRs is the serial resistance and capacitors' ESR. This resistance in the power delivery network creates a voltage droop between the VR output and the CMP supply input. Modern CPU VR designs use Adaptive Voltage Positioning (AVP) to control this effect [18] [19]. Splitting the PDN into 16 individual power delivery networks, assuming that the total capacity is unchanged, results with 16X higher resistance in the power delivery to each core. On symmetric workloads, each core would consume 1/16 of the total current compared to a single shared VR, with 16X higher resistance, resulting in the same voltage droop. On asymmetric workloads, however, more power budget is allocated to one core; as a result, the higher power core consumes more current and suffers a higher voltage droop. This effect is modeled into the power delivery network of this study.

3.3 Power Model

The power model is based on lab measurements of a real product, a 2.6 GHz Intel® Core™ 2 Duo platform. A standard PC board was instrumented with thermal control head, replacing the industry standard heat sink. A 0.1mOhem serial shunt resistors

connected in series to the CPU core and I/O voltage regulators and a FLUK 2645A used to record supply voltage and current at 100mSec intervals and calculate power consumption. ACPI interface [21] has been used to modify voltage and frequency and to measure voltage, frequency and power scaling curves of DVFS. The measured part provided DVFS capability from 800MHz to 2.6GHz and voltage from 0.85V to 1.2V. Frequency only can go further down to 100Mz. Leakage has been measured as a function of voltage, at a controlled junction temperature equal to the maximum specification (100'C). Measured leakage of the tested part was 30% of total power. The model assumes that the active cores will run at max power all the time and therefore leakage will be equal to the measured value. Hot spots and non-uniform power distribution effects on leakage are ignored. A cycle accurate simulator of the Intel® Core™ 2 Duo with power modeling has been used to evaluate interconnect power resulting 10% out of the total power.

Power does not scale very well with the advent of process technology. Therefore, it is assumed that future high power CMPs will have to be designed such that the nominal rated frequency and power of all cores will be at the minimum operating voltage that is allowed by the process. DVFS control will increase the frequency whenever power headroom is available. The model in this study performs DVFS up to the nominal power and not exceeds it. It is also possible to reduce the frequency below the nominal point without changing voltage and achieve linear power reduction.

This study does not rely on predicting absolute values of voltage, frequency and power of future technologies. Instead, relative values are employed: The above measured voltage, frequency, power and leakage at the minimum voltage are defined as 100% and DVFS or DFS are described as percentage of the reference point. DVFS up to 200% of the nominal frequency and DFS down to 50% are considered, as noted in Figure 2.

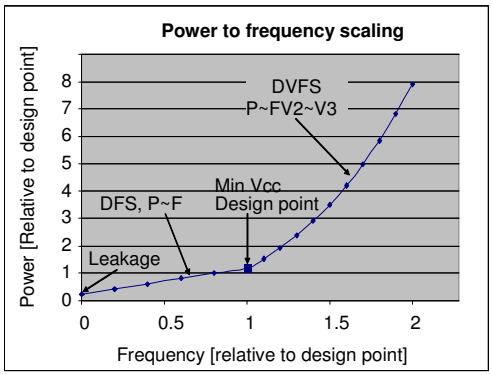

Figure 2: DVFS and DFS model

**In the cubic range, V and F are scaled together. In the linear range, only F is scaled.*

3.4 Benchmark construction and framework

This study evaluates performance in power-constrained conditions. CMP power consumption and maximum power delivery capabilities affect the CMP in thermally significant time intervals which range from a few milliseconds to many seconds. Running a few seconds workload of a multi-gigahertz CPU on a cycle accurate simulator is not practical. However, some detailed information about workload characteristics (described below) can be achieved only on a cycle accurate simulator. On the other hand, running benchmarks on a real product provides good and reliable results. Thus, a hybrid approach has been used: benchmark measurements have been collected on a 2.6 GHz Intel® Core™ 2 Duo in a PC platform and on a cycle accurate simulator, and a full model consisting of performance scores and scaling as well as DVFS modeling was constructed.

3.4.1 Benchmark simulation and testing

A set of 26 components of SPEC-2000 has been used for this study. Table 1 demonstrates the computed benchmark data.

Table 1: Benchmark Parameters

SPEC int	Scaled Power	Perf. Scaling with freq.	FIFO impact
gzip	48%	0.95	0.13%
vpr	44%	0.68	2.92%
gcc	35%	0.67	0.92%
mcf	49%	0.30	2.92%
crafty	33%	0.99	0.59%
parser	60%	0.78	1.29%
eon	42%	0.99	0.00%
perlbmk	50%	1.00	0.31%
gap	45%	0.56	1.14%
vortex	60%	0.73	1.45%
bzip2	49%	0.70	0.71%
twolf	97%	0.99	4.68%
Int_rate	51%	0.77	1.42%

SPEC FP	Scaled Power	Perf. Scaling with freq.	FIFO impact
wupwise	51%	0.23	1.09%
swim	83%	0.00	1.84%
mgrid	54%	0.06	0.89%
applu	57%	0.13	0.46%
mesa	47%	0.86	0.00%
galgel	100%	0.56	0.66%
art	79%	0.23	1.21%
equake	37%	0.08	1.84%
facerec	53%	0.00	0.53%
ammp	66%	1.00	1.66%
lucas	55%	0.05	0.97%
fma3d	59%	0.37	1.06%
sixtrack	40%	0.98	0.03%
apsi	79%	0.65	0.49%
fp_rate	62%	0.09	0.91%

The cycle accurate simulator enabled evaluating the impact of the added latency of the FIFO synchronizers on each workload for various latencies and clock frequencies. In addition, each benchmark was executed on the real CPU at different frequencies, measuring CPU power and performance scores. The scaled power column represents the average power of each benchmark expressed as a percentage of the highest power application (GALGEL). It is assumed that the nominal operating point is intended to run GALGEL on all 16 cores simultaneously. Whenever a lower power application is executed, there is power headroom which is used to increase frequency and provide higher performance.

Performance scaling with frequency in the table expresses the ratio of benchmark score change to frequency change. The FACEREC benchmark for example is completely constrained by memory and therefore changing frequency does not affect its performance, resulting in a ratio of 0. AMMP and PERLBMK on the other hand, are compute-bound and therefore scale perfectly with frequency, represented by a ratio of 1. A linear dependency over the entire frequency range is assumed, and has been shown a reasonable assumption on this micro-architecture.

3.4.2 Constructing multi core workloads using Monte-Carlo modeling

How should workloads be assigned to the different cores of the CMP? Previous studies carefully selected a handful of representative applications [1] but these are too specific and are applicable only to a small number of cores. In this study a Monte-Carlo simulation approach is employed in order to evaluate and present a wide span of possibilities for multi-core CMP workloads. For each run, a set of up to 16 applications out of the 26 SPEC components is randomly selected and the same set is applied to all the different topologies and policies, thus neutralizing workload effect on the results. This procedure is repeated for 200 runs, covering the range of ±2σ of workload distribution. Some studies, however, evaluate average results and therefore 50 runs are sufficient. Two types of studies are performed, fully loaded CMP running 16 threads simultaneously and partially loaded CMP running randomly picked workloads of pre-defined number of threads, fewer than 16.

The study is aimed to maximize the performance within the set of constraints. Relative values are compared to a nominal baseline. A total relative performance is defined as the sum of the individual relative benchmark results, normalized to the number of active cores, $\frac{1}{n}\sum_{i=1}^{n} Perf_i$

3.5 Power Management Policies

For each of the topologies described in Section 3.1, the study seeks the best power management policy that maximizes performance under the power constraints of that particular topology. Managing the cores consists of distributing the threads to the different cores and performing DVFS according to the power management policy. Efficient power management policies require some knowledge of the applications being executed. It has been shown in prior work that it is possible to evaluate the power of the CPU [3] [4] and the workload scalability [7] [8] at run time. Thus, the investigated power management policies use the power and scalability information collected on the Intel® Core™ 2 Duo as input to the power management algorithms. A single policy is used for the each of the benchmarks for the entire run, based on average characteristics; adaptive policies are beyond the scope of this paper.

The first modeling step evaluates a fully loaded CMP running 16 threads using each of the different topologies. For each topology, an oracle approach is employed first to figure out the best possible performance for the topology. The oracle has been implemented using the Generalized Reduced Gradient (GRG2) algorithm [22] solving for maximum performance under all constraints (maximum and minimum frequency, total power and maximum current). Subsequently, various policies are implemented and compared to the oracle. The following policies have been implemented:

- Either power or scalability is selected as the input parameter (X axis in Figure 3).
- The selected parameter is used as an input argument for the evaluated frequency scaling functions described in Figure 3. The output of the function is operation frequency
- Frequency and voltage are scaled together, according to the measured model described in Figure 2.

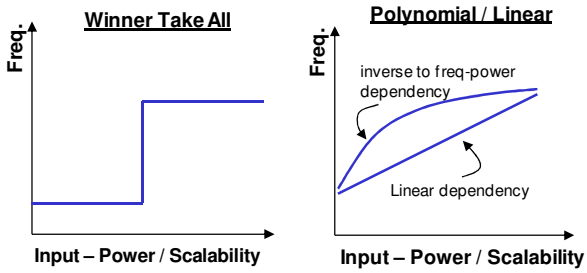

Figure 3: Scaling functions

The Winner Take All (or Binary) policy is based on the leftmost function of Figure 3. A fixed threshold is selected, and the input parameter is compared to the threshold. If the parameter is higher than the threshold the frequency and voltage are increased to maximum. If not, the frequency and voltage remain low. This policy has been applied with the parameters of scalability and power, where the latter is applied both to the highest power applications (aka positive power) and to the lowest power applications (aka negative power). Several different threshold values have been studied.

The polynomial policy (rightmost chart of Figure 3) scales frequency and voltage of each core based on the inverse function of P~f(F), namely third root down to minimal voltage and subsequently linear.

The linear policy also shown in Figure 3 is similar to the polynomial policy above with linear dependency of frequency to the input parameters, namely scalability, positive power and negative power.

A fourth scaling function assigns random frequencies (and the corresponding voltages) to the cores based on a "first come first served" policy.

The rationale behind these policies is based on the expectation of higher benefit when higher frequencies are assigned to the cores that can gain the most from it (high scalability) or that are not high power and therefore could extract more performance from

the same amount of power. Random assignment is given as a reference.

In the multiple voltage domains topology, the resulting frequency of each core is evaluated and if the resulting current exceeds the current capability defined for that particular run, the frequency is bounded by the maximum allowed current. In the single voltage domain topology, the same test is performed for the single shared power delivery. Finally, all values are normalized such that the total CPU power meets the defined power constraint and the power delivery limitations.

In this study, the order at which workloads are assigned to different cores does not affect the model and the resulting performance. In practice, physical proximity of hot spots does impact junction temperature limitation but thermal modeling lies beyond the scope of this paper.

The next evaluation step models partial threads workload. 2,4,8,12,14 and 16 threads have been studied. In a single voltage domain the power is shared and can provide for fewer active cores. However, routing power to only parts of the chip creates asymmetry, and this effect has been incorporated into the voltage-frequency models. On a multiple voltage domain topology, each individual power distribution network is required to meet the current constraint. The entire study is repeated on the partial threads model, using the oracle.

3.6 Clustered Topology

Last, a set of clustered topologies is investigated. The cores are grouped into N=2,4,8 clusters consisting of 8,4,2 cores respectively. Clustering obtains topologies that cover the ground between the two extremes, a single domain and 16 separate domains. Clustering allows sharing power delivery between several cores while maintaining the capability to scale cluster frequency asymmetrically. The model has been adapted to reflect N power delivery networks and clock domains. The entire study is repeated on the clustered model, using the oracle.

4. Detailed Studies and Results

4.1 Baseline performance

The baseline for all other comparisons is based on the single voltage domain, single clock domain topology. For each set of randomly selected applications, frequency is increased using DVFS to fully utilize the power headroom, as long as maximum frequency and power delivery limits are not exceeded. Each performance score is divided by the score of the CMP running the same workload at the nominal frequency. 100% means that the total score of the run with DVFS scaling is equal to the baseline score, either because there has been no power headroom to increase frequency or because the applications has not gain from the frequency speedup. 125% means that the total score of the run with DVFS scaling is 25% higher performance than the baseline. The results are described in Figure 4.

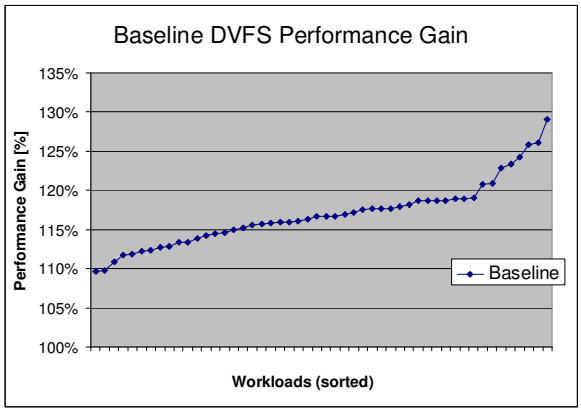

Figure 4: Baseline performance on 200 runs of randomly selected benchmarks

The horizontal axis lists the 200 runs. The chart shows the relative performance speedup of these runs, compared to the baseline, sorted in ascending order. Observe that there is up to 27% performance gain with average of 17% that can be achieved by utilizing the power headroom. The rest of the studies in this paper use these results as a base line, attempting to extract more performance than the simple single frequency scheme by utilizing more complicated topologies and policies.

4.2 Fully threaded workload – Oracle

A multiple voltage domain topology with single and multiple clock domains are evaluated and compared to the baseline of Sect. 4.1. The oracle provides the optimal setting possible for each workload on each topology. 200 random workloads are modeled in the Monte-Carlo simulator. For each individual workload, the performance ratio between each pair of topologies is calculated and the ratios are sorted low to high. Results for power delivery with 150% headroom are presented in Figure 5 relative to the baseline in Figure 4. The following terminology is employed in the charts: 1V indicates a single voltage domain, nV indicates multiple voltage domains, 1C and nC indicate single and multiple clock domains, respectively.

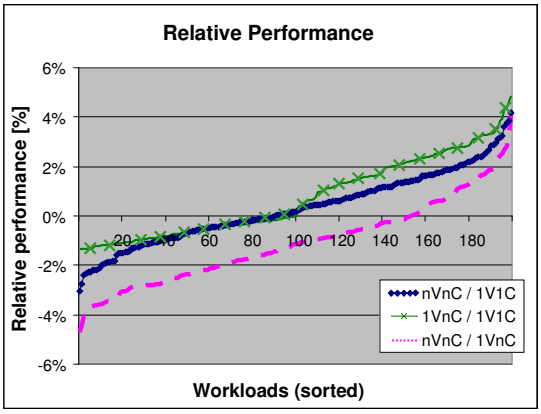

Figure 5: Performance relative to the baseline with 150% power headroom

The single voltage domain with single clock domain is a special case of the single voltage domain with multiple clock domains, with all frequencies equal (1VnC/1V1C). The chart shows that 100 out of the 200 workloads gain from selecting different frequencies for some of the cores. The other 100 workloads have equal frequency on both topologies. The performance degradation of 0% to 1.3% on these 100 workloads on multiple clock domains topology is due to the cost of FIFO synchronization. Comparing multiple voltage and clock domains to a single voltage and clock domain (nVnC/1V1C) shows that 50% of the workloads gain performance from splitting the voltage domains while the other 50% lose. The overall average gain is +0.3%. This result is in contrast with previous studies that ignored power delivery and topology costs.

Comparing multiple voltage and clock domains to a single voltage and multiple clock domains (nVnC/1VnC) shows that 78% of the workloads lose performance from splitting the voltage domains with an overall average loss of 1.1%.

This study has shown that multiple independent voltage and frequency domains may not always and unconditionally be a winning topology as previously claimed. Power delivery constraints shift the balance more towards a single voltage domain while higher power delivery capabilities shift it in favor of split voltage domains.

4.3 Fully threaded workload – Policies

Which policy is most suitable for each topology? The results are shown in Table 2.

Table 2: Matching policies to topologies

1VnC		
	Max	Average
WTA 50%	5.84%	1.3%
WTA 33%	4.41%	0.6%
WTA 10%	1.23%	0.0%
WTA by Power 50%	22.76%	6.9%
Linear by SCA	9.60%	6.1%
Linear by power	49.76%	36.6%
Polinomial by SCA	5.23%	3.3%
Random	33.28%	19.9%

nVnC		
	Max	Average
WTA 50%	2.90%	0.8%
WTA 33%	3.37%	0.8%
WTA 10%	4.63%	1.7%
WTA by Power 50%	4.60%	2.3%
Linear by SCA	2.72%	1.5%
Linear by power	5.77%	3.8%
Polinomial by SCA	3.58%	1.5%
Random	8.66%	4.3%

WTA = Winner Take All, SCA = Scalability

The values in the table indicate distances from the respective oracle. Lower numbers are better, meaning closer to the maximum performance reached by the oracle. For all topologies, the policies that provide the best results are based on Scalability (SCA). This result is explained by the fact that assigning the highest frequency to the workloads that can gain the most performance out of it yield the best overall performance. For a single voltage domain topology, WTA policy with low SCA threshold provides the best performance while on multiple voltage domains linear dependency of frequency on scalability or higher threshold for the WTA policy are the best. For a single clock domain, obviously there is only one policy that scales the single frequency up until meeting the most constraining parameter (power, power delivery or maximum frequency). Interestingly, random assignment of workloads to the cores (which can represent a first come first served policy) yield poor results; this caused by the polynomial frequency-to-power relation. We conclude that utilizing the potential of performance headroom requires knowledge about the workload and implementation of a power management policy that makes good use of such knowledge.

4.4 Partially threaded workloads

So far all 16 cores were used. Workloads that occupy only part of the cores have much higher power and thermal headroom. In the extreme case of a single threaded application, the entire power budget can be assigned to the single core running this single thread. If no other constraint exists, this core can run at 16X power of a single core. This may not be possible due to power delivery and maximum voltage and frequency constraints. In the following, all topologies are evaluated with workloads that activate partial number of cores. An oracle figures out the best performance that can be achieved at each of the combinations. The study is repeated for various power delivery constraints ranging from 130% to 250% of the nominal power delivery requirement. The two extreme cases are shown in Figure 6.

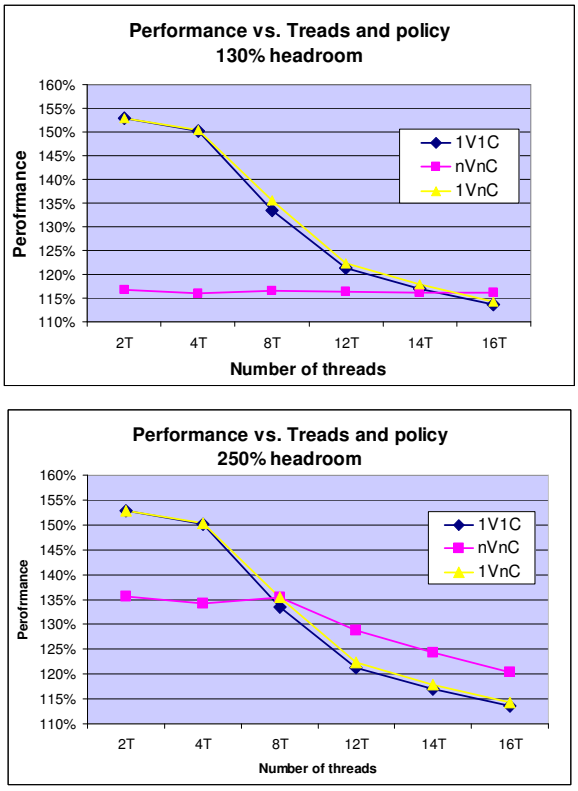

Figure 6: Partial threads

The charts show average performance gain over baseline of nominal frequency. Single voltage domain is not impacted by power delivery headroom because sharing current among cores delivers sufficient current and therefore the frequency and performance are constrained by maximum frequency and power before they reach maximum current limits. On multiple voltage domains, however, each core is constrained by its own individual power delivery and therefore it becomes the dominant limiter, especially in workloads with a small number of threads. The higher power delivery capability, the higher the performance each individual core can achieve. As a result, there is a crossover point at which multiple power domains deliver higher average performance than a single voltage domain. With 250% power delivery headroom, multiple voltage domains provide better performance for workloads with 12 threads or more. The charts show average performance of all randomly selected runs. Examining the individual workloads shows distribution of results. The 16 threads case is described in details in Figure 5 with some applications gaining performance and others losing performance on the multiple voltage domain topology.

4.5 Clustered topology

The previous sections indicate that multiple voltage and clock domains have the potential of providing higher performance than the single voltage and clock domain. As described above, splitting the CMP into multiple voltage domains comes at the cost of degraded power delivery networks, becoming especially constraining in single or low threaded applications, whereas heavily threaded workloads with enough power delivery headroom gain from multiple voltage domains. Looking to benefit from both worlds, the CMP can be partitioned not to individual cores but rather into clusters, each consisting of several cores that share the same voltage domain. The clustered CMP architecture may also partition other architectural elements such as the shared cache and the interconnect; however, this study focuses on power so that such other architectural opportunities are left unexplored.

This study assigns threads to clusters in a round robin fashion - first one thread for each cluster, than second thread for each cluster until all threads has been assigned a core. This thread distribution policy assures that the power consumption of the different threads will be distributed between the power domains. As above, various power delivery capabilities are modeled. An example of 8 clusters. is shown in Figure 7. More run results are described in Table 3

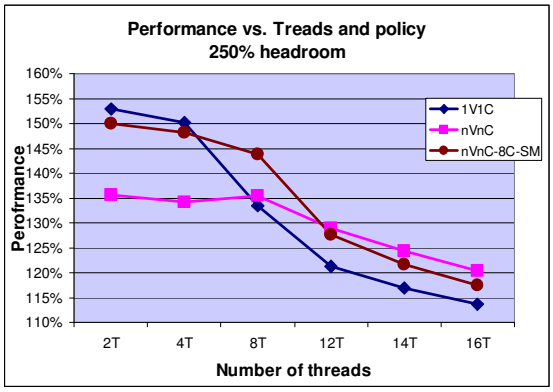

Figure 7: Clustered CMP

Clustered approach achieves the best of all worlds. In a single thread application it allows sharing power delivery and provides performance similar to performance that can be achieved by a single voltage domain. In workloads that run many threads, it provides the benefits of multiple clock and voltage domains. In the crossing point, (8T in this example), clustering outperforms both single and multiple voltage domains. On CMP with more cores, benefit of clustering is expected to be higher.

4.6 Optimal cluster size

The following analysis seeks a rule for optimizing cluster size. Clustered CMP modeling has been performed on various parameters and the minimum quadratic distance from the best possible scenario has been calculated (Table 3).

Table 3: Optimal cluster size

	110%	130%	150%	200%	250%
1V1C	7.1%	11.4%	13.2%	14.8%	16.6%
1VnC	5.1%	9.0%	10.7%	12.4%	14.1%
nVnC-2C	28.6%	13.0%	14.1%	15.4%	17.5%
nVnC-4C	45.8%	14.7%	13.3%	12.2%	13.9%
nVnC-8C	55.6%	21.9%	16.5%	9.8%	7.6%

The two best scenarios for each cluster size are highlighted, demonstrating a clear "diagonal" behavior, indicating correlation between power delivery headroom and the number of clusters or cluster size. For constrained power delivery of 110% the best performance is achieved by a single cluster (a single voltage domain) with all 16 cores sharing the same power delivery. 130% headroom is best served by 2-4 clusters, 150% by 4-8 clusters and 200% and above with 8 clusters.

5. Conclusions

In this paper we studied the effects of multiple voltage domains and multiple clock domains on CMP power and performance. A realistic CMP model was employed; real constraints related to voltage regulation, power delivery networks and power requirements were implemented and their impact was examined in respect to the overall power and performance.

Adding few pragmatic constrains to the analysis, yield unexpected results; unlike previous understanding that splitting power and clock domains is unequivocally beneficial in terms of power and performance, we found out that in reality, it depends on the number of cores, workload characteristics, power delivery architecture, frequency and current constraints.

A novel experimental methodology for CMP analysis was described, combining cycle accurate simulation and execution on actual sample machines. Monte-Carlo simulation of random subsets of the SPEC 2000 benchmark enabled neutralizing benchmarking bias. A reliable cubic/linear frequency-power model was introduced. Oracles were devised for creating baseline expectations. Topologies and policies were created to effectively explore the design space.

Three approaches were contemplated and studied: a fully threaded workload that occupies all cores, a partial threading for investigating power diversion from unoccupied to busy cores, and a clustered micro-architecture inspired by the difficulties of

providing a large number of different power supplies and power delivery networks on- and off-chip.

Studies using oracle showed the best performance that can be achieved in each given scenario. We then addressed the question of how a power management unit may control the individual cores and achieve this best possible performance. We evaluated power and scalability characteristics as an input to a control algorithm, and we compared various control algorithms in order to achieve performance that is as close as possible to the oracle results.

The study results show that power delivery is a primary constraint that limits the benefit of independent voltage and frequency domains. In unconstrained environment, individually controlling each core achieves higher performance by assigning higher power budget to the cores that benefit the most from it. This asymmetry however requires cores capable of operating at their worst case conditions. The probability of all cores working simultaneously at worst case conditions is very low. Sharing physical resources among cores therefore benefits the majority of workloads. Separate voltage and frequency domain provide on average 6% higher performance compared to a single power and clock domain when all cores are active. On a single thread workload however it loses 14%. The performance lost is higher if the power delivery is more constrained. Clustering the cores into groups allows individual frequency scaling together with physical resources sharing among cores. Results show the existence of optimal cluster size as a function of physical characteristics of the CPU.

In order to fully utilize the potential performance, we need to know workload characteristics and implement an algorithm that controls the cores based on this knowledge. Random assignment of frequencies to cores, or using a wrong parameter, is worse than no individual control at all. Scalability proved as best parameter to construct a high performance algorithm. . Knowing which core benefits the most from increased frequency and assigning that core a higher power budget maximizes the overall performance. There is no single algorithm that fits all topologies and constraints but in general greedy algorithms (WTA) based on workload scalability perform well in all scenarios as long as the threshold point is tuned to meet the topology.

6. REFERENCES

[1] Isci, C., Buyuktosunoglu A., Cher C., Bose, P., Martonosi, M. An Analysis of Efficient Multi-Core Global Power Management Policies: Maximizing Performance for a Given Power Budget. In *Proceedings of 39th Annual IEEE/ACM International Symposium on Microarchitecture*, 2006.

[2] Ogras, U. Y., Marculescu, R., Choudhary, P., and Marculescu, D. Voltage-frequency island partitioning for GALS-based networks-on-chip. In *Proceedings of the 44th Annual Design Automation Conference,* June 2007, San Diego, California.

[3] Bellosa, F., The benefits of event driven energy accounting in power-sensitive systems, In *Proceedings of the 9th workshop on ACM SIGOPS European workshop: beyond the PC: new challenges for the operating system*, September 2000, Kolding, Denmark.

[4] Contreras, G., Martonosi, M. Power prediction for Intel XScale® processors using performance monitoring unit events. In *Proceedings of the 2005 international symposium on Low power electronics and design*, August 2005, San Diego, CA, USA

[5] Kihwan C., Soma, R., Pedram, M. Fine-grained dynamic voltage and frequency scaling for precise energy and performance tradeoff based on the ratio of off-chip access to on-chip computation times. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 24(1), 18-28. January 2005.

[6] Hsu, C., Feng, W. Effective dynamic voltage scaling through CPU-boundedness detection. Proc. In *Proceedings of 4th Workshop Power-Aware Computer Systems*, December 2004.

[7] Wu, Q., Martonosi, M., Clark, D.W., Reddi, V.J., Connors, D., Wu, Y., Lee, J., Brooks, D. Dynamic-Compiler-Driven Control for Microprocessor Energy and Performance. *Micro, IEEE*, 26(1), 119-129. January 2006.

[8] Hill, M.D., Marty, M.R. Amdahl's Law in the Multicore Era. *Computer* , 41(7),33-38. July 2008.

[9] Annavaram, M., Grochowski, E., Shen, J. Mitigating Amdahl's law through EPI throttling. In *Proceedings of the 32nd International Symposium on Computer Architecture*, June 2005.

[10] Grochowski, E., Ronen, R,, Shen, J., Wang. H. Best of both latency and throughput. In *Proceedings of the International Conference on Computer Design*, 236–243. October 2004, San Jose, CA.

[11] Joseph, R., Brooks, D., Martonosi, M. Control techniques to eliminate voltage emergencies in high performance processors. In *Proceedings of the 9th International Symposium on High-Performance Computer Architecture*, February 2003.

[12] Gupta, M. S., Oatley, J. L., Joseph, R., Wei, G.-Y., Brooks, D. Understanding voltage variations in chip multiprocessors using a distributed power-delivery network. In *Proceedings of Design, Automation, and Test in Europe Conference*, April 2007.

[13] Kim, W., Gupta M., Wei, G. Y., Brooks, D. System level analysis of fast, per-core DVFS using on-chip switching regulators. In *Proceedings of International Symposium on High-Performance Computer Architecture*, February 2008.

[14] Naveh, A., Rotem, E., Mendelson, A., Gochman, S., Chabukswar, R., Krishnan, K., Kumar, A. Power and Thermal Management in the Intel Core Duo Processor. *Intel Technology Journal*. 10(2). 109-122. May 2006.

[15] Liang, X,, Brooks, D., Wei G.-Y. A process-variation-tolerant floating-point unit with voltage interpolation and variable latency. In *Proceedings of International Solid-State Circuits Conference*, February 2008.

[16] Zane, B., Karen, R., IA Product Update, In Cebit, March 2009. http://www.intel.com/corporate/pressroom/emea/deu/cebit/pdfs/CeBIT_Client_Tech_Briefing.pdf

[17] Intersil, Document Number: FN9289.3, 14 February, 2007, http://www.intersil.com/data/fn/FN9289.pdf

[18] Waizman, E., Chung, C. Y. Resonant free power network design using extended adaptive voltage positioning (EAVP)

methodology. *IEEE Trans. Adv. Package*. 24(3). 236–244, August 2001.

[19] Voltage Regulator-Down (VRD) 11.0 Processor Power Delivery Design Guidelines For Desktop LGA775 Socket, November 2006. http://www.intel.com/assets/pdf/designguide/313214.pdf

[20] Ren, Y., Yao, K., Xu, M., Lee, F.C. Analysis of the power delivery path from the 12 V VR to the microprocessor. In *Proceedings of the Applied Power Electronics Conference and Exposition*, 285-291. 2004

[21] Advanced Configuration and Power Interface (ACPI) Specification, October 2006. http://www.acpi.info/

[22] Schittkowski, K., Zillober, C., Zotemantel, R. Numerical comparison of nonlinear programming algorithms for structural optimization, Structural Optimization, 7(1), 1–19. 1994.

[23] Gochman, S., Mendelson, A., Naveh, A., Rotem, E. Introduction to Intel® Core™ Duo processor architecture. *Intel Technology Journal,* 10(2), 2006.