Report from Dagstuhl Seminar 16052

Dark Silicon: From Embedded to HPC Systems

Edited by

Hans Michael Gerndt[1], Michael Glaß[2], Sri Parameswaran[3], and Barry L. Rountree[4]

1 **TU München, DE,** gerndt@in.tum.de
2 **FAU Erlangen-Nürnberg, DE,** michael.glass@fau.de
3 **UNSW – Sydney, AU,** sridevan@cse.unsw.edu.au
4 **LLNL – Livermore, US,** rountree@llnl.gov

Abstract

Semiconductor industry is hitting the utilization wall and puts focus on parallel and heterogeneous many-core architectures. While continuous technological scaling enables the high integration of 100s–1000s of cores and, thus, enormous processing capabilities, the resulting power consumption per area (the power density) increases in an unsustainable way. With this density, the problem of Dark Silicon will become prevalent in future technology nodes: It will be infeasible to operate all on-chip components at full performance at the same time due to the thermal constraints (peak temperature, spatial and temporal thermal gradients etc.). However, this is not only an emerging threat for SoC and MPSoC designers, HPC faces a similar problem as well: The power supplied by the energy companies as well as the cooling capacity does not allow to run the entire machine at highest performance anymore. The goal of Dagstuhl Seminar 16052 "Dark Silicon: From Embedded to HPC Systems" was to increase the awareness of the research communities of those similarities and to work and explore common solutions based on more flexible thermal/power/resource management techniques both for runtime, design time as well as hybrid solutions.

Seminar January 31 to Febuary 3, 2016 – http://www.dagstuhl.de/16052

1998 ACM Subject Classification C.3 [Special-Purpose and Application-based Systems] Real-time and embedded systems, C.1.4 [Processor Architectures] Parallel Architectures

Keywords and phrases dark silicon, embedded, hpc, parallel computing, performance analysis and tuning, power density, power modelling, programming tools, resource management

Digital Object Identifier 10.4230/DagRep.6.1.224

Edited in cooperation with Michael Glaß

1 Executive Summary

Hans Michael Gerndt
Michael Glaß
Sri Parameswaran
Barry L. Rountree

Topic

Dark Silicon

Semiconductor industry is hitting the utilization wall and puts focus on parallel and heterogeneous many-core architectures. While continuous technological scaling enables the high

Dark Silicon: From Embedded to HPC Systems, *Dagstuhl Reports*, Vol. 6, Issue 1, pp. 224–244
Editors: Hans Michael Gerndt, Michael Glaß, Sri Parameswaran, and Barry L. Rountree
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

integration of 100s-1000s of cores and, thus, enormous processing capabilities, the resulting power consumption per area (the power density) increases in an unsustainable way. With this density, the problem of Dark Silicon will become prevalent in future technology nodes: It will be infeasible to operate all on-chip components at full performance at the same time due to the thermal constraints (peak temperature, spatial and temporal thermal gradients etc.).

Recent research work on power management for Dark Silicon aims at efficiently utilizing the TDP (Thermal Design Power) budget to maximize the performance or to allocate full power budget for boosting single-application performance by running a single core at the maximum voltage or multiple cores at nominal level for a very short time period. Control-based frameworks are proposed to find the optimal trade-off between power and performance of many-core systems under a given power budget. The controllers are coordinated to throttle down the power when the system exceeds the TDP and to assign the task to the most suitable core to get the optimal performance. The work on near-threshold computing (NTC) enables operating multiple cores at a voltage close to the threshold voltage. Though this approach favors applications with thread-level parallelism at low power, it severely suffers from errors or inefficiency due to process variations and voltage fluctuations. On the other hand, the computational sprinting approach leverages Dark Silicon to power-on many extra cores for a very short time period (100s of millisecond) to facilitate sub-second bursts of parallel computations through multi-threading but thereby wasting a significant amount of energy due to leakage current. When doing so, it consumes power that significantly exceeds the sustainable TDP budget. Therefore, these cores are subsequently power-gated after the computational sprint. Alternate methods are Intel's Turbo Boost and AMD's Turbo CORE technologies that leverage the temperature headroom to favor high-ILP applications by increasing the voltage/frequency of a core while power-gating other cores. These techniques violate the TDP constraint for a short period (typically in terms of 10s of seconds) until the critical temperature is reached and then switches to a nominal operation. However, in case of dependent workloads, boosting of one core may throttle the other due to thermal coupling (i.e. heat exchange between different cores sharing the same die). Therefore, these boosting techniques lack efficiency in case dependent tasks of an application mapped to two different cores or, in general, for multiple concurrently executing applications with distinctive/dependent workloads.

State-of-the-art boosting techniques assume a chip with only 10-20 cores (typically 16) and accordingly a full chip temperature violation for short time. However, in a large-scale system (with 100s–1000s cores), temperature hot spots may occur on certain chip portions far before the full chip's average temperature exceeds the critical temperature. Therefore, a chip may either get damaged before reaching the full chip critical temperature or TDP needs to be pessimistically designed. Advanced power management techniques are required to overcome these challenges in large-scale environments.

HPC – Dark Power

The energy consumption of HPC systems is steadily growing. The costs for energy in the five year lifetime of large scale supercomputers already almost equal the cost of the machine. It is a necessity to carefully tune systems, infrastructure and applications to reduce the overall energy consumption. In addition, the computing centers running very big systems face the problem of limited power provided by the energy providers and of the requirement for an almost constant power draw from the grid. The big machines, especially future exascale systems, are able to use more power if they are run at highest performance of all components than can be provided by the energy company. Thus, a carefully optimized power distribution is necessary to make most efficient use of the provided power. The second aspect is the requirement of an almost constant power draw: Sudden changes from 20 MW to 10 MW for example, will be dangerous for the components of the power grid. In addition, the contracts with the energy companies force the centers use the same power all the time by charging more, if it drops below or exceeds certain limits. These challenges also require a careful and flexible power and resource management for HPC systems.

For a certain class of high-end supercomputer, there is a standard pattern of power consumption: During burn-in (and perhaps while getting a result to go onto the top-500 list) the machine will run dozens or hundreds of instances of Linpack. This code is quite simple and often hand-optimized, resulting in an unusually well-balanced execution that manages to keep vector units, cache lines and DRAM busy simultaneously. The percent of allocated power often reaches 95 % or greater, with one instance in recent memory exceeding 100 % and blowing circuit breakers. After these initial runs, however, the mission-critical simulation codes begin to execute and they rarely exceed 60 % of allocated power. The remaining 40 % of electrical capacity is dark: just as unused and just as inaccessible as dark silicon. While we would like to increase the power consumption (and thus performance) of these simulation codes, a more realistic solution in the exascale timeframe is hardware overprovisioning. This solution requires buying more compute resources than can be executed at maximum power draw simultaneously. For example, if most codes are expected to use 50 % of allocated power, the optimal cluster would have twice as many nodes.

Making this a feasible design requires management of power as a first-class resource at the level of the scheduler, the run-time system, and on individual nodes. Hardware power capping must be present. Given this, we can theoretically move power within and across jobs, using all allocated power to maximize throughput. The purpose of this seminar is to find this optimal level.

Hybrid (Design-time & Run-time) Resource Management

Today's complex applications need to exploit the available parallelism and heterogeneity of – non-darkened – cores to meet their functional and non-functional requirements and to gain performance improvements. From a resource management's point of view, modern many-core systems come with significant challenges: (a) Highly dynamic usage scenarios as already observable in today's "smart devices" result in a varying number of applications with different characteristics that are running concurrently at different points in time on the system. (b) Due to the constraints imposed by the power density, the frequency at which cores can be operated as well as their availability as a whole, are subject to change. Thus, resource management techniques are required that enable a resource assignment to applications that satisfies their requirements but at the same time can consider the challenging dynamics of modern many-cores as a result of Dark Silicon.

Traditional techniques to provide a binding or pinning of applications to processor that are optimal and predictable with respect to performance, timing, energy consumption, etc. are typically applied at design time and result in a kind of static system design. Such a static design may, on the one hand, be too optimistic by assuming that all assigned resources are always available or it may require for a kind of over-allocation of cores to compensate for worst-case scenarios, e.g., a frequent unavailability of cores due to Dark Silicon. Hence, the dynamic effects imposed in Dark Silicon require for novel modeling techniques already at design time.

Approaches that focus on pure run-time resource management are typically designed with flexibility in mind and should inherently be able to dynamically react to changing applications as well as to the described effects of Dark Silicon. But, future run-time resource management should not only react to a possible violation of a maximum power-density constraint, but also be able to proactively avoid such situations. The latter is an important aspect of the system's dependability as well. At the same time, such dynamic resource management is also required to regard the applications' requirements. Here, a careful consideration on whether pure run-time management strategies enable the amount of predictability of execution qualities required by some applications becomes necessary.

A recent research direction focuses on hybrid (design-time and run-time) approaches that explore this field of tension between a high predictability of design-time approaches and the dynamic adaptivity of run-time resource management. In such approaches, design-time analysis and optimization of the individual applications is carried out to capture information like core allocation, task binding, or message routing and predict resulting quality numbers like timeliness, energy consumption, or throughput. This information is then passed to the run-time resource management that then dynamically selects between the pre-optimized application embeddings. Such strategies may not only be able to achieve application requirements even in such highly dynamic scenarios, but could even balance the requirements of the individual applications with the system's requirements – in particular the maximum power density. On the other hand, coarse-grained resource management as required for core allocation etc. may be considered to happen on a longer time scale. The effects of Dark Silicon are instead on a smaller time scale with temperature almost immediately following changing workloads, thus, requiring for an intervention of the resource-management infrastructure. Therefore, novel concepts are required that enable a fine-grained resource management in the presence of Dark Silicon – both in the context of abstraction layer and time scale – without sacrificing the required efficiency but also predictable realization of application requirements via coarse-grained resource management.

Goals

Traditionally, resource management techniques play an important role in both domains – targeting very different systems. But, as outlined before, resource management may be the key to tackle the problem of dark silicon that both communities face. The aim of this seminar is to give an overview of the state of the art in the area of both embedded and HPC. It will make both groups aware of similarities and differences. Here, the competences, experiences, and existing solutions of both communities shall stimulate discussions and co-operations that hopefully manifest in innovative research directions for many-core resource management in the dark silicon era.

Overview of Contributions

This seminar presentations on the state-of-the-art in power and energy management in HPC and on techniques mitigating the Dark Silicon problem in embedded systems. In a joint session commonalities and differences as well as collaboration potential in the area of Dark Silicon were explored. This subsection gives an overview of the topics covered by the individual speakers in the seminar. Please refer to the included abstracts to learn more about individual presentations.

The HPC-related presentations where started with an overview presentation by Barry Rountree from the Lawrence Livermore National Laboratory. He introduced the field of HPC and of exascale systems. The new challenge is that these systems will be power limited and the hardware is overprovisioned. Techniques increasing the efficient usage of the available power need to be developed. Exascale systems will be heterogeneous, even systems with homogeneous cores become heterogeneous due to production variability which takes effect under power limits. Careful distribution of power among jobs and within jobs as well as application and system configurations for jobs will be important techniques for these power limited and overprovisioned systems.

Axel Auweter added to this introduction deep insights into the electricity market in Germany, its complex price structure, and the challenges for German compute centers to act successfully on that market.

An introduction from the embedded field to Dark Silicon was given by Sri Parameswaran from the University of New South Wales. The continuous decrease in feature size without an appropriate decrease in the threshold voltage leads to increased power density. Between 50 % and 90 % of dark silicon is expected in future chips. Mitigation techniques are energy reduction techniques as well as spatial and temporal dimming of cores. Considerable energy reduction can be achieved from heterogeneity on various levels, e.g., heterogeneous cores and the DarkNoC approach.

Dark Silicon due to Power Density

Several techniques were presented to mitigate the effect of power density. Santiago Pagani presented *spatial and temporal dimming of cores* to make best use of the thermal distribution on the chip. He and Andrey Semin talked also about *boosting* the core frequency to exceed the power limit for a short time period to speedup computation. Sergio Bampi presented *near threshold computing* as a potential solution based on further lowering the threshold voltage. Michael Niemier explored the potential of *new transistor technology* to mitigate the Dark Silicon effect.

Dark Silicon due to Limited Power

Mitigation techniques in this field are quite similar in mobile computing and HPC, although the overall objective is a bit different. While in mobile computing the minimal power required to meet the QoS requirements of applications is the goal, in HPC it is to go as fast as possible with the available power, may be considering energy efficiency and system throughput as well.

The following approaches relevant for mobile computing and HPC were presented: *Heterogeneity* in various hardware aspects can be used to reduce the energy consumption of computations. Siddarth Garg and Tulika Mitra covered *performance heterogeneity* in scheduling tasks for big/little core combinations. Tulika Mitra and Andrea Bartolini talked about using *function heterogeneity*, e.g. accelerators, in mobile computing and HPC to increase energy efficiency. The *Heterogeneous Tile Architecture* was introduced in the presentations of Sri Parameswaran and Santiago Pagani as a general architecture enabling exploitation of heterogeneity to mitigate the Dark Silicon effect.

Another approach is to determine the most efficient *application and system configuration. Static tuning* of parameters, such as the power budget of an application, were presented by Michael Knobloch and Tapasya Patki. *Dynamic tuning* techniques were covered in the presentations of Michael Gerndt, Martin Schulz, and Per Gunnar Kjeldsberg. Jonathan Eastep introduced the GEO run-time infrastructure for distributed machine-learning based power and performance management.

Kirk Cameron highlighted the unexpected effects of changing the core frequency due to non-linear dependencies. Jürgen Teich talked about *Invasive Computing* providing dynamic resource management not only for improving certain non-functional application aspects but also for increasing the predictability of those aspects.

Wolfgang Nagel and Sri Parameswaran presented *energy efficient network architectures*. They covered heterogeneous on-chip network architectures and wireless communication within compute clusters.

Approximate computing was presented by Sergio Bampi. It allows trading off accuracy and energy. Pietro Cicotti covered in his presentation *data movement optimization* within a CPU to save energy.

Application and system monitoring is a pre-requisite for many of the above techniques. Michael Knobloch, Wolfgang Nagel, and Kathleen Shoga presented application and system monitoring techniques based on software as well as hardware instrumentation. Many compute centers are installing infrastructures to gather sensor values from the whole facility to enable future analysis. In addition to performance and energy measurements for application, higher level information about the application characteristics is useful in taking tuning decisions. Tapasay Patki presented *application workflows* as a mean to gather such information.

Besides these generally applicable techniques, some presentations covered also techniques that are specific to HPC installations with their batch processing approach and large compute systems.

Andrea Bartolini highlighted in his presentation the holistic multiscale aspect of power-limited HPC. The application, the compute system, and the *cooling infrastructure* have to be seen as a complex integrated system. *Power-aware scheduling*, presented by Tapasya Patki and Andrea Bartolini, can significantly improve the throughput of power-limit HPC systems and *moldable jobs* can improve the effect of power-aware scheduling significantly. Isaias Compres presented *Invasive MPI*, an extension of MPI for programming moldable application.

Conclusion

At the end of the seminar a list of takeaway messages was collected based on working-group discussions followed by an extensive discussion of all participants:

1. Dark silicon is a thermal problem in embedded and a power problem in HPC. HPC can cool down while in the embedded world you can't. Therefore HPC can power up everything if they have enough power. But the costs for providing enough power for rare use cases have to be rectified.
2. Better tools are required on both sides to understand and optimize applications.
3. Better support for optimizations is required through the whole stack from high level languages down to the hardware.
4. In both communities run-time systems will get more important. Applications will have to be written in a way that run-time systems can work effectively.
5. Task migration is of interest to both groups in combination with appropriate run-time management techniques.
6. Embedded also looks at specialized hardware designs while HPC has to use COTS. In HPC, the machine architecture might be tailored towards the application areas. Centers are specialized for certain customers.
7. Heterogeneity on architecture level is important to both groups for energy reduction.

8. Better analyzable programming models are required, providing composable performance models.
9. HPC will have to live with variability. The whole tuning step has to change since reproducibility will no longer be given.
10. Hardware-software co-design will get more important for both groups.
11. Both areas will see accelerator-rich architectures. Some silicon has to be switched off anyway, thus these can be accelerators that might not be useful for the current applications.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Beyond Power Capping – Coping with the Complexity of the German Electricity Market

Axel Auweter (LRZ – München, DE)

Germany is one of the countries with the highest costs for electricity in the world. On top, the regulations and pricing scheme is overly complex. Yet, the availability of power saving and capping techniques, intelligent resource management and power consumption prediction models in HPC opens up the possibility for leveraging this complexity in smart ways. This presentation explains the pricing of electricity in Germany and how current and future developments from the EEHPC domain might help optimize the TCO for German HPC centers.

3.2 Dark Power: Applying Lesson from Dark Silicon to Power-Constrained High Performance Computing

Barry Rountree (LLNL – Livermore, US)

The field of high performance computing is experiencing a sea change: where previous machines were limited by the number of compute nodes that could be purchased, future exascale machines will be primarily limited by the amount of power than can be brought into the center. The US Department of Energy has a target for the first exaflop machine to consume no more than 20 megawatts. Compared to early petaflop machines, this effectively means a 1000 x increase in performance for an 3x increase in power.

This change calls into question how we think about performance. If power is going to be the limiting factor, the we should be maximizing its utilization. Current HPC codes make use of only 60 % of allocated power. Scaling these codes up naively to exascale and beyond implies that 40 % of the electrical infrastructure would remain idle for most of the lifetime of the machine. In short, we need a new model for machine design and evaluation that uses all available power to maximize job performance and system throughput.

3.3 Multiscale Energy-Thermal Management for Green Supercomputers

Andrea Bartolini (University of Bologna, IT & ETH Zürich, CH)

In the last decade large high performance computing systems as well as processing elements have become power and energy limited. At system scale energy provisioning and cooling power and facility design limit the available power budget for each machine installation while at component scale the end of Dennard's scaling makes the power consumption the limiting factor for the performance of the computing devices. Today's processors performances are thermally and power limited, while today's supercomputers performance are power, cooling and cost limited. In this talk I will present a set of tools, methodology and research results on the evaluation of the impact of temperature on the energy-efficiency of the supercomputer and internal components and opportunities for advanced and holistic management of thermally constrained large scale computing systems.

3.4 The Law of Unintended Consequences for Dark Silicon

Kirk W. Cameron (Virginia Polytechnic Institute – Blacksburg, US)

Joint work of Hung-Ching Chang, Bo Li, Godmar Back, Ali Raza Butt, Kirk W. Cameron
Main reference H.-C. Chang, B. Li, G. Back, A. R. Butt, K. W. Cameron, "LUC: Limiting the Unintended Consequences of Power Scaling on Parallel Transaction-Oriented Workloads", in Proc. of the 2015 IEEE Int'l Parallel and Distributed Processing Symp. (IPDPS'15), pp. 324–333, IEEE, 2015.
URL http://dx.doi.org/10.1109/IPDPS.2015.99

In 1936, Harvard University sociologist Robert Morton wrote a paper entitled "The unanticipated consequences of purposive social action", where he described how government policies often result in both positive and negative unintended consequences. The lesson from Morton's work was that unexpected consequences in complex social systems, at the time relegated to theology or chance, should be evaluated scientifically.

Since the performance effects of dark silicon management are largely unknown, these potentially valuable features introduce risk and uncertainty in complex, large-scale, high-performance systems. While dark silicon promises to address power and energy limitations for emergent systems, Morton teaches us that relegating performance behavior to chance is just as likely to result in negative consequences. For example, there is mounting evidence that when processors are fixed at fully-powered, highest frequency (i.e., disabling dynamic frequency scaling), performance can worsen. Thus, challenges and opportunities abound for dark silicon to be adopted by the HPC community.

In this presentation, I will demonstrate that "faster is NOT always better" when managing power and performance of large scale systems. In essence, slowing down CPU frequency (or powering down system components) can speed up performance as much as 50 % for some I/O intensive applications. I show that identifying the root cause of such slowdowns is wrought with challenges. I will describe how modeling and runtime systems can limit these anomalies but that dark silicon will undoubtedly lead to more unintended consequences for high-performance systems.

References

1 *LUC: Limiting the Unintended Consequences of Power Scaling on Parallel Transaction-Oriented Workloads.* IPDPS 2015: 324-333, Hyderabad, India, 2015.

3.5 Data Movement in Dark Silicon Systems

Pietro Cicotti (San Diego Supercomputer Center, US)

Joint work of Laura Carrington, Pietro Cicotti

As power limitations induce the presence of dark silicon, a new dimension appears in the configuration and optimization space of applications and systems. In order to optimize performance and efficiency, execution must be combined with an understanding of available resources (darkened and not) and potential gain/cost tradeoffs in using them.

A fundamental aspect of this optimization problem is associated with the need to move data. Leveraging dark silicon implies that data must be moved to powered resources, and then fetched back. For example, in computational sprinting and invasive computing, data must move to the claimed resources and then be flushed back when the resources are released. In addition, with the ability to finely select and configure the resources claimed, it is important to correctly estimate and select an optimal configuration of resources.

In this context, carefully tuning systems and applications requires understanding data access patterns and the effect of different memory hierarchies. In this presentation, I will discuss our work in modeling memory and creating tools to analyze and eventually manage data movement dynamically.

3.6 Elastic Execution Models and Energy Aware Job Scheduling in HPC

Isaias Alberto Compres Urena (TU München, DE)

Power density has increased in recent computing Integrated Circuits (IC), while computer hardware designs have maintained largely the same heat dissipation properties. This situation has led to Dark Silicon scenarios, where large parts of an IC must remain powered off to keep it in safe operating levels. Parallels can be drawn in HPC systems, where power limits require nodes or partitions to be turned off. Elastic execution models allow distributed memory applications to adapt to changes in resources. Job schedulers can manipulate such applications individually to achieve global energy requirements such as power level stabilization.

3.7 An Introduction to GEO: A New Open Source Extensible Power Management Framework from Intel

Jonathan Eastep (Intel – Hillsboro, US)

URL http://geopm.github.io/geopm

In this talk, I will provide an intro to GEO (Global Energy Optimization). GEO is an open source, scalable, extensible runtime system and power management framework for HPC systems from Intel. It provides out-of-the-box power management technology to mitigate application load imbalance by redistributing power to the application's critical path, and it provides extensibility to new power management strategies through a plug-in architecture. A goal of the project is to provide a convenient platform that HPC power researchers can build their research on and accelerate innovation in HPC power management. The runtime as well as plug-ins are licensed with a permissive BSD license to encourage community and industry adoption and collaboration. See geopm.github.io/geopm for more project information and a link to the source code.

3.8 Scheduling for Dark Silicon Servers

Siddharth Garg (New York University, US)

Joint work of Siddharth Garg, Umit Ogras

Heterogeneous processors with multiple core types, for example, the so-called "big-little" processors, are becoming increasingly common-place. This talk will focus on energy and thermally-aware scheduling for heterogeneous servers; in particular, we will discuss a family of so-called "threshold" policies that preferentially assign jobs to power efficient cores and utilize larger cores only when the number of outstanding jobs exceeds a threshold. We will also discuss policies for parallelizable jobs.

3.9 Energy Efficiency Tuning: From Autotune to READEX

Hans Michael Gerndt (TU München, DE)

The European AutoTune project developed the Periscope Tuning Framework (http://periscope.in.tum.de) for pre-production tuning of HPC applications [1]. Tuning plugins capture expert knowledge for a certain tuning aspect and search for optimal tuning parameter settings. Periscope provides a rich framework to tuning plugins, e.g., standard search algorithms, performance analysis services, static program information, and automatic experiment executions. Plugins use expert knowledge to structure the search process and to reduce the search space. Tuning plugins can use performance analysis information to, for example, determine the size distribution of messages and using this information to restrict the range of values for the eager threshold of the MPI library. Standard search algorithms are used to finally generate scenarios that are experimentally evaluated. The scenario is specified by the plugin and the real execution and measurement of the objective function is done automatically by Periscope.

The focus of automatic tuning in the AutoTune project was on design time tuning. The best setting for the tuning parameters is determined before production runs of the application. This best setting is then static for the execution. In the new Horizon 2020 project READEX (http://www.readex.eu) this work is extended for runtime tuning [2]. It follows the approach of scenario-based optimization from embedded systems. At design time the application is analyzed and a tuning model is constructed that captures the best configurations for various runtime situations. This tuning model is passed to the runtime and dynamic switching happens between configuration if new runtime situations are encountered. The Periscope Tuning Framework is used for the design time analysis and Score-P is extended with the READEX Runtime Library for dynamic configuration switching.

References

1 Michael Gerndt, Eduardo César and Siegfried Benkner (Eds.). *Automatic Tuning of HPC Applications – The Periscope Tuning Framework*. Shaker Verlag, ISBN 978-3-8440-3517-9, 2015

2 Y. Oleynik, M. Gerndt, J. Schuchart, P. G. Kjeldsberg, W. E. Nagel. *Run-Time Exploitation of Application Dynamism for Energy-Efficient Exascale Computing (READEX)*. IEEE 18th International Conference on Computational Science and Engineering (CSE), pp. 347–350, 2015

3.10 Scenario based design of dynamic embedded applications

Per Gunnar Kjeldsberg (NTNU – Trondheim, NO)

Joint work of Special Interest Group on Scenario Driven Design for Embedded Systems
URL http://www.es.ele.tue.nl/scenarios/

System scenario methodologies propose the use of different scenarios, e.g., different heterogeneous platform configurations, in order to exploit variations in computational and memory needs during the lifetime of an application. The system scenario methodology consists of a design-time and a run-time stage. The application is analyzed at design-time and different execution paths and variations in processing and memory demands are identified. Situations with similar N-dimensional Pareto cost requirements are grouped into a limited number of scenarios. At run-time the current situation is detected, and the platform is reconfigured accordingly, e.g., through remapping of tasks on processors, voltage and frequency scaling, changing power modes of memories, turning on and off processing cores and accelerators, etc. Compared with use-case scenarios, system scenarios exploit detailed knowledge of the application, giving rise to much larger performance and energy gains.

This talk will present the system scenario design methodology including results from implementation examples in the embedded systems domain.

3.11 Energy-efficient HPC – A Tools Perspective

Michael Knobloch (Jülich Supercomputing Centre, DE)

Main reference R. Schöne, J. Treibig, M. F. Dolz, C. Guillen, C. B. Navarrete, M. Knobloch, B. Rountree, "Tools and methods for measuring and tuning the energy efficiency of HPC systems", Scientific Programming, 22(4):273–283, 2014.
URL http://dx.doi.org/10.3233/SPR-140393

Energy consumption of applications and power draw of large-scale installations has become a major topic in HPC on the road to Exascale. A detailed analysis of power and energy consumption is necessary in order to understand system and application characteristics and control them for maximum efficiency. However, traditional performance analysis tools face multiple challenges obtaining power and energy relevant data. In this talk I present the work done by JSC and its partners in multiple energy-efficiency related projects and discuss the requirements on hardware in order to improve power and energy consumption analysis.

3.12 The impact of new transistor technologies on core scaling trends (and dark silicon)

Michael Niemier

Joint work of Michael Niemier, Robert Perricone, X. Sharon Hu, Joseph Nahas

Continued transistor scaling no longer yields exponential performance gains due in part to the growth of dark silicon (DS). Both industrial and government sponsors are actively pursuing the development of new transistor technologies that may re-enable voltage scaling, offer I-V characteristics that lead to simpler and/or more efficient circuits when compared to CMOS functional equivalents, etc.

This talk will discuss architectural-level modeling efforts that build upon the framework developed in [1, 2], which originally considered DS in the context of core scaling efforts via the PARSEC benchmark suite [3]. The Notre Dame group has worked to unify architectural-level benchmarking [1, 2] with device-level benchmarking [4, 5] (that considers devices being studied under the umbrellas of various SRC and DARPA initiatives) to provide insight as to how voltage scaling could impact the viability of core scaling – and hence the spread of DS.

Interestingly, for high thermal design power TDPs (125 W), projections suggest that low voltage devices achieve a speedup of just 2X on average in the best case when compared to 15 nm high performance (HP) CMOS. Moreover, per [5] the 15 nm CMOS datapoint is representative of 2018 technology – which will undoubtedly come to market well-before emerging low voltage devices (which may presently exist in simulation only). Not surprisingly, for lower TDPs (5W), emerging low voltage devices fare much better when compared to HP CMOS – and speedups of approximately 10X appear possible. However, speedups of approximately 2.5X are projected over 2018 low power (LP) CMOS. In addition to these results, other benchmarking data and possible paths forward will be highlighted.

References

1 H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, “Dark silicon and the end of multicore scaling,” in Computer Architecture (ISCA), 2011 38th Annual International Symposium on, June 2011, pp. 365–376.

2 H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, “Power challenges may end the multicore era,” Commun. ACM, vol. 56, no. 2, pp. 93102, Feb. 2013.

3 C. Bienia, S. Kumar, J. P. Singh, and K. Li, “The parsec benchmark suite: Characterization and architectural implications,” in Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques, ser. PACT’08. New York, NY, USA: ACM, 2008, pp. 72–81.

4 D. Nikonov and I. Young, “Overview of beyond-cmos devices and a uniform methodology for their benchmarking,” Proceedings of the IEEE, vol. 101, no. 12, pp. 2498–2533, 2013.

5 D. Nikonov and I. Young, “Benchmarking of beyond-cmos exploratory devices for logic integrated circuits,” Exploratory Solid-State Computational Devices and Circuits, IEEE Journal on, vol. 1, pp. 3–11, Dec 2015.

3.13 Improving Energy-Efficiency through Heterogeneity in Mobile Platforms

Tulika Mitra (National University of Singapore, SG)

Joint work of Tulika Mitra, Thannirmalai Somu Muthukaruppan, Anuj Pathania, Alok Prakash, Mihai Pricopi, Vanchinathan Venkataramani, Sanjay Vishin
Main reference T. S. Muthukaruppan, A. Pathania, T. Mitra, "Price theory based power management for heterogeneous multi-cores", in Proc. of the 19th ACM Int'l Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS'14), pp. 161–176, ACM, 2014.
URL http://dx.doi.org/10.1145/2541940.2541974

The impact of dark silicon is more pronounced in the mobile platforms due to the absence of active cooling in these systems. In order to cope with the effect of dark silicon, mobile system-on-chip designs embrace heterogeneous multi-core architectures where cores with different functional characteristics (CPU, GPU, DSP, non-programmable accelerators) and/or power-performance characteristics (simple versus complex micro-architecture) co-exist on the same die. Given an application, only the cores that best fit the application can be switched on leading to faster and energy efficient computing. We present application-aware, software-level runtime management strategies to leverage the potential of heterogeneous multi-core architectures.

3.14 Workflow Analysis – A map between Applications and System Resource Needs

David Montoya (Los Alamos National Lab., US)

Workflow has always been used to describe jobs and applications progressing and interacting with systems. As systems become more tightly integrated with varied architectures and with varied feedback loops added to better balance resource utilization, do we have a map that includes the application? When you envision the overall HPC environment being made up of applications and system components that are made up of workflows interacting with each other, it becomes apparent that we don't have the tools to assess this interaction.

In this presentation I will describe an effort at LANL where we have started by developing a workflow taxonomy with layers describing the application stack, how we have used it for initial assessment for future machine needs, and the potential to further define lower layers to integrate with mapping of machine layers of workflow. As this evolves it brings in application and system performance collection, deriving workflow performance, and system monitoring as key initial capabilities.

3.15 Performance, Energy, Structure, and Materials: What we have to learn, and how we will address the Challenges!

Wolfgang E. Nagel (TU Dresden, DE)

Parallelism on chips and technology improvements nowadays have led to dark silicon, power budgets, and temperature and energy variations, which heavily depend on hardware features and usage profiles of the applications. Running thousands of these sockets in parallel lead to reasonable challenges not only in the field of load balancing, but also in the energy usage. The talk describes research work of the Dresden group in the field of energy measurement on the microsecond level, embedded in the collaborative research center HAEC (highly adaptive energy-efficient computing). In HAEC, research technologies are developed to enable computing systems with high energy-efficiency without compromising on high performance. As part of that, a novel concept (HAEC Box) of how computers can be built by utilizing innovative ideas of optical and wireless chip-to-chip communication is explored. This CRC is embedded in the excellence cluster cfAED (Center for Advanced Electronics Dresden) where also material research is done to shape the time after CMOS. The talk will describe the general approach and the implications on programming challenges in future systems.

3.16 Mitigating the Power Density and Temperature Problems in the Nano-Era

Santiago Pagani (KIT – Karlsruher Institut für Technologie, DE)

Joint work of Santiago Pagani, Heba Khdr, Waqaas Munawar, Dennis Gnad, Muhammad Shafique, Siddharth Garg, Minming Li, Jian-Jia Chen, Jörg Henkel
Main reference S. Pagani, H. Khdr, W. Munawar, J.-J. Chen, M. Shafique, M. Li, J. Henkel, "TSP: Thermal Safe Power – Efficient power budgeting for many-core systems in dark silicon", in Proc. of the 2014 Int'l Conf. on Hardware/Software Codesign and System Synthesis (CODES'14), Art. 10, ACM, 2014.
URL http://dx.doi.org/10.1145/2656075.2656103

In this talk we introduce the Dark Silicon problem and discuss our research efforts for mitigating the associated power density and temperature issues. Specifically, we talk about mapping/patterning and how making smart mapping decisions can reduce the peak temperature on the chip. We introduce the Thermal Safe Power (TSP) concept for efficient power budgeting, in which the power budget depends on the number of active cores. We present some of our experiments comparing single/constant frequency solutions against boosting techniques. Finally, we discuss MatEx, an efficient analytical transient and peak temperature computation tool. In conclusion, power and performance efficiency should be jointly optimized at multiple hardware and software layers of the system stack. If all this is considered, there is a good chance that the Dark Silicon problem can be avoided.

3.17 System-Wide Power Management in High-Performance Computing

Tapasya Patki (LLNL – Livermore, US)

Main reference T. Patki, D. K. Lowenthal, A. Sasidharan, M. Maiterth, B. Rountree, M. Schulz, B. R. de Supinski, "Practical Resource Management in Power-Constrained, High Performance Computing", in Proc. of the 24th Int'l Symposium on High-Performance Parallel and Distributed Computing (HPDC'15), pp. 121–132, ACM, 2015.
URL http://dx.doi.org/10.1145/2749246.2749262

One of the key challenges on the path to exascale supercomputing is power management. Supercomputing centers today are designed to be worst-case power provisioned, leading to two main problems: limited application performance and under-utilization of procured power. This talk will introduce hardware overprovisioning: a power-efficient design approach for future supercomputing centers that addresses the aforementioned problems. Power-aware resource management policies targeted toward overprovisioned HPC systems will also be discussed.

3.18 System Software for Power Limited HPC Systems: Challenges and Solutions

Martin Schulz (LLNL – Livermore, US)

Power and energy consumption are critical design factors for any next generation large-scale HPC system. The costs for energy are shifting the budgets from investment to operating costs, and more and more often the size of systems will be determined by its power needs. As a consequence, it is likely that we will end up with power limited systems that can no longer power all their components at peak power. In these systems, system software must manage power caps at all layers of the system to ensure only the available power is used in the system and that this available power is used efficiently. In this talk, I will discuss the need and opportunities of power-limited systems and the challenges they pose and present system software techniques at different levels of the software stack that can help users successfully and efficiently exploit power-limited systems. In particular, I will present an approach to mitigate processor manufacturing variability at large scale [1], a runtime system – Conductor [2] – to steer power within an MPI application in order to maximize the utilization of an available power budget, and an operating system component – PowSched [3] – to exploit unused power resources in power constrained HPC systems.

References

1 Y. Inadomi, T. Patki, K. Inoue, M. Aoyagi, B. Rountree, M. Schulz, D. Lowenthal, Y. Wada, K. Fukazawa, M. Ueda, M. Kondo, and I. Miyoshi. Analyzing and mitigating the impact of manufacturing variability in power-constrained supercomputing. In *Proce. of the Int'l Conf. for High Performance Computing, Networking, Storage and Analysis*, SC'15, pp. 78:1–78:12, USA, 2015. ACM.

2 A. Marathe, P. E. Bailey, D. K. Lowenthal, B. Rountree, M. Schulz, and B. R. Supinski. *High Performance Computing: 30th Int'l Confe., ISC High Performance 2015, Frank-*

furt, Germany, July 12-16, 2015, Proceedings, chapter A Run-Time System for Power-Constrained HPC Applications, pp. 394–408. Springer Int'l Publishing, Cham, 2015.

3 D. A. Ellsworth, A. D. Malony, B. Rountree, and M. Schulz. Dynamic power sharing for higher job throughput. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC'15, pp. 80:1–80:11, USA, 2015. ACM.

3.19 Processor power and performance variability and impact on applications performance

Andrey Semin (Intel GmbH – Feldkirchen, DE)

Modern Intel microprocessors contain embedded controller called PCU (power control unit) that is in full control of processor execution state and mode of operation. Many power management and performance-related features are implemented with the use of this controller logic. At the same time some of PCU control operations presents challenges for the performance and parallel scaling of the HPC applications, specifically in the area of performance reproducibility. One of the specific challenges is that PCU makes CPU frequency dependent on consumed power, while we note that power is dependent on frequency, voltage, as well as temperature of the circuit. The observed variability in performance and power is noted by many HPC system users, and these observations are summarized in this presentation. In the conclusion we propose an "uncertainty principle" that governs power and frequency (or the cycle time) variability dependencies.

3.20 Livermore Computing Monitoring Infrastructure

Kathleen Sumiko Shoga (LLNL – Livermore, US)

Power, energy, and thermal constraints require us to make smarter use of our resources to get the best performance. There are, however, many factors that go into the performance of applications run in a large computing center. At Livermore Computing, we are deploying monitoring across the center in a multi-level fashion from the facilities level down to the hardware performance counter level. Gathering and analyzing this data will enable us to make better choices when it comes to tradeoffs for resources and future system designs.

3.21 Introduction to Dark Silicon – Problems and Techniques

Sri Parameswaran

Joint work of Muhammad Shafique, Haseeb Bokhari, Joerg Henkel, Sri Parameswaran

In this talk we discuss aspects of dark silicon, starting with the definition of dark silicon, its origins and the reason why modern chips are becoming larger, yet are unable to be powered on completely. This problem is exacerbated in embedded systems, where cooling is limited and thus only a fraction of the chip can be turned on at any one time. The second part of the talk explains some of the methods used to overcome the issues arising from dark silicon, and explains some of the opportunities afforded to designers and users of modern chips. Finally, we delve deeper in to one of the methods used to mitigate the problems of dark silicon and talk about creating a NoC which utilizes the area afforded by dark silicon to improve reliability and energy efficiency.

3.22 Adaptive Restriction and Isolation for Increasing *-Predictability

Jürgen Teich (Universität Erlangen-Nürnberg, DE)

Resource sharing and interferences of multiple threads of one, but even worse between multiple application programs running concurrently on a Multi-Processor System-on-a-Chip (MPSoC) today make it very hard to provide any timing or throughput-critical applications with time bounds. Additional interferences result from the interaction of OS functions such as thread multiplexing and scheduling as well as complex resource (e.g., cache) reservation protocols used heavily today. Finally, dynamic power and temperature management on a chip might also throttle down processor speed at arbitrary times leading to additional varations and jitter in execution time. This may be intolerable for many safety-critical applications such as medical imaging or automotive driver assistance systems.

Static solutions to provide the required isolation by allocating distinct resources to safety-critical applications may not be feasible for reasons of cost and due to the lack of efficiency and unflexibility.

In this Dagstuhl presentation, we first review definitions of predictability. We distinguish two techniques for improving predictability called restriction and isolation and present new definitions for predictability. Subsequently, new techniques for adaptive isolation of resources including processor, I/O, memory as well as communication resources on demand on an MPSoC are introduced based on the paradigm of Invasive Computing. In Invasive Computing, a programmer may specify bounds on the execution quality of a program or even segment of a program followed by an invade command that returns a constellation of exclusive resources called a claim that is subsequently used in a by-default non-shared way until being released again by the invader. Through this principle, it becomes possible to isolate applications automatically and in an on-demand manner. In invasive computing, isolation is supported on all levels of hardware and software including an invasive OS. Together with restriction (of input uncertainty), the level of on-demand predictability of program execution qualities may be fundamentally increased.

For a broad class of streaming applications, and a particular demonstration based on a complex object detection application algorithm chain taken from robot vision, we show how jitter-minimized implementations become possible, even for statically unknown arrivals of other concurrent applications.

Participants

- Axel Auweter
LRZ – München, DE
- Sergio Bampi
Federal University of Rio Grande do Sul, BR
- Andrea Bartolini
University of Bologna, IT & ETH Zürich, CH
- Kirk W. Cameron
Virginia Polytechnic Institute – Blacksburg, US
- Pietro Cicotti
San Diego Supercomputer Center, US
- Isaias Alberto Compres Urena
TU München, DE
- Jonathan Eastep
Intel – Hillsboro, US
- Siddharth Garg
New York University, US
- Hans Michael Gerndt
TU München, DE
- Michael Glaß
Universität Erlangen-Nürnberg, DE
- Per Gunnar Kjeldsberg
NTNU – Trondheim, NO
- Michael Knobloch
Jülich Supercomputing Centre, DE
- Tulika Mitra
National University of Singapore, SG
- David Montoya
Los Alamos National Lab., US
- Wolfgang E. Nagel
TU Dresden, DE
- Michael Niemier
University of Notre Dame, US
- Santiago Pagani
KIT – Karlsruher Institut für Technologie, DE
- Sri Parameswaran
UNSW – Sydney, AU
- Tapasya Patki
LLNL – Livermore, US
- Barry L. Rountree
LLNL – Livermore, US
- Martin Schulz
LLNL – Livermore, US
- Andrey Semin
Intel GmbH – Feldkirchen, DE
- Kathleen Shoga
LLNL – Livermore, US
- Jürgen Teich
Universität Erlangen-Nürnberg, DE