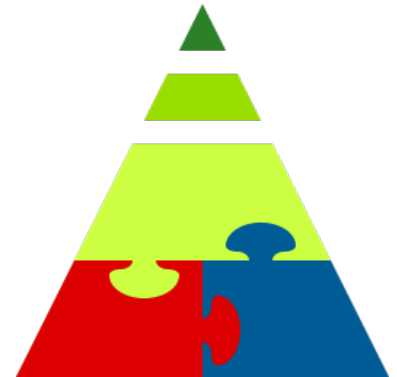


chist-era

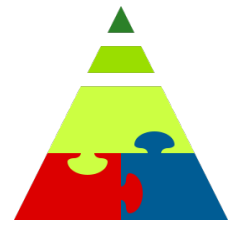


DIVIDEND

Heterogeneous Vertically-Integrated Data Centers

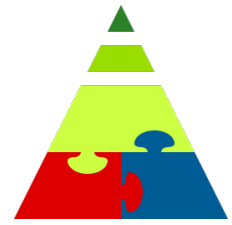
Babak Falsafi

The future of IT is Data



- Data growth (by 2015) = 100x in ten years [IDC 2012]
 - Population growth = 10% in ten years
- Monetizing data for commerce, health, science, services, ...
- Big Data is shaping IT & pretty much whatever we do!

Data Shaping All Science & Technology

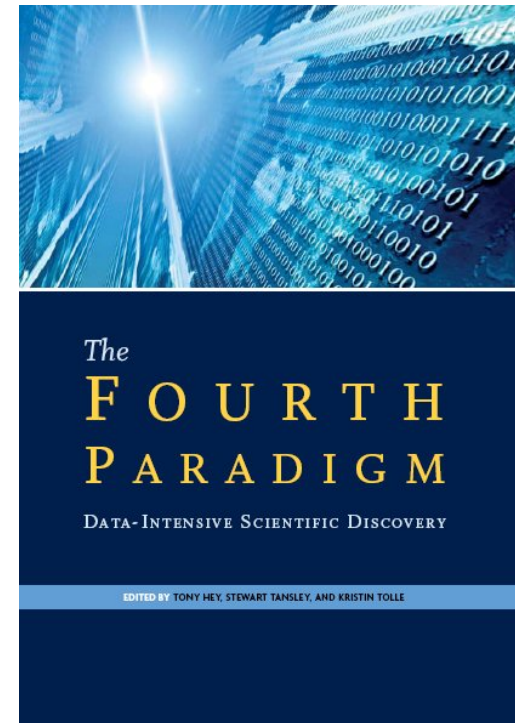


Science entering 4th paradigm

- Analytics using IT on

- Instrument data
- Simulation data
- Sensor data
- Human data
- ...

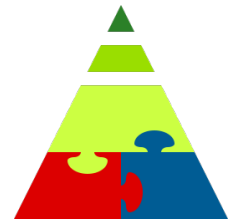
Complements theory, empirical science & simulation



Data-centric science key for innovation-based economies!

Source: James Hamilton, 2014

mvdirona.com/jrh/TalksAndPapers/JamesHamilton_Reinvent20131115.pdf



Perspective on Scaling



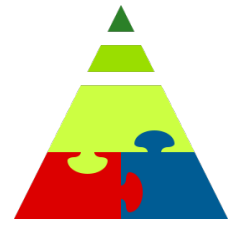
Every day, AWS adds enough new server capacity to support all of Amazon's global infrastructure when it was a \$7B annual revenue enterprise

AWS
re:Invent

Daily IT growth in 2014 = All of AWS in 2004!

Warning!

Datcenters are not Supercomputers



- Run heterogeneous data services at massive scale
- Driven for commercial use
- Fundamentally different design, operation, reliability, TCO
 - Density 10-25KW/rack as compared to 25-90KW/rack
 - Tier 3 (~2 hrs/downtime) vs. Tier I (upto 1 day/downtime)
 - and lots more

Datcenters are the IT utility plants of the future

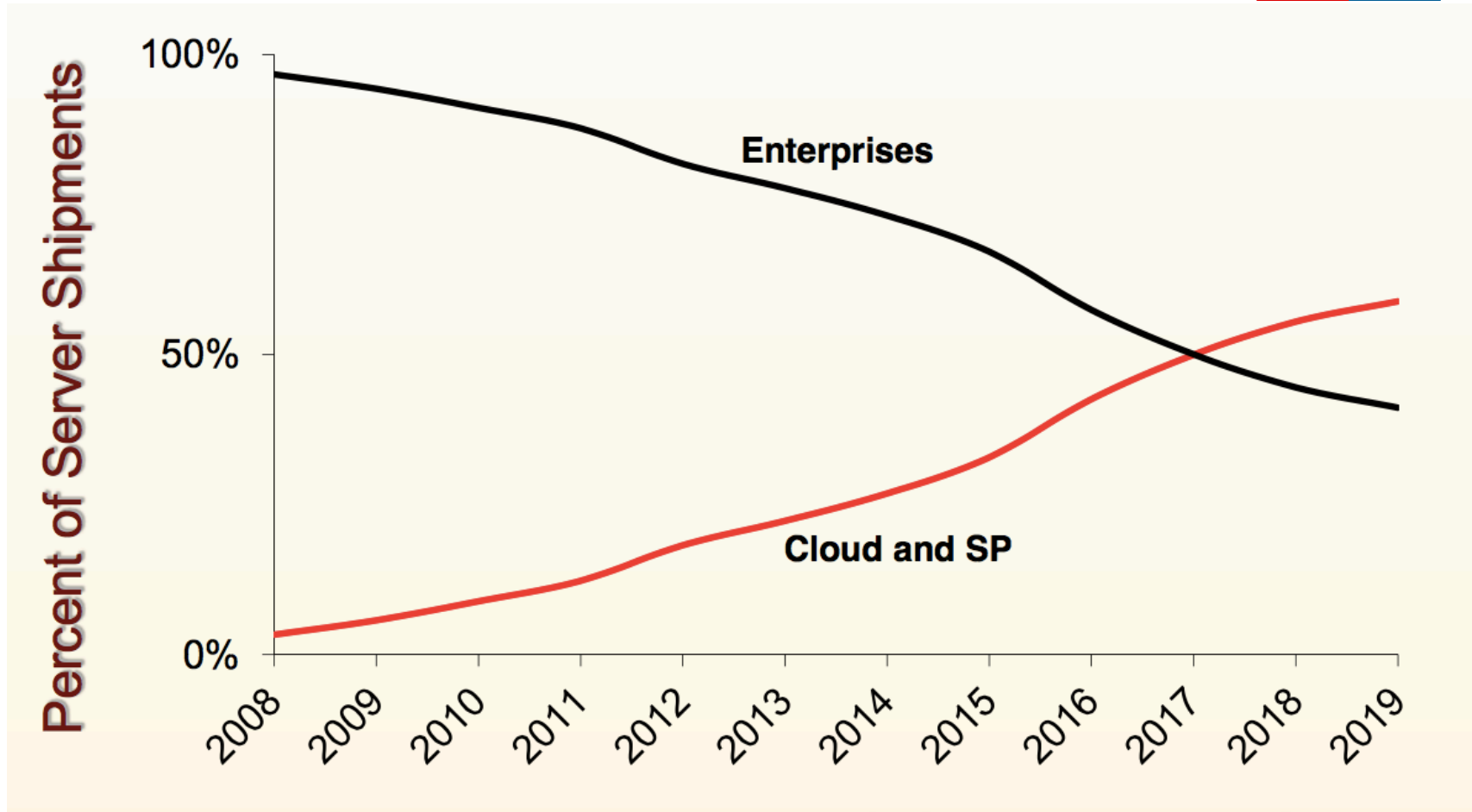
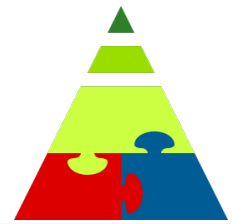


Supercomputing



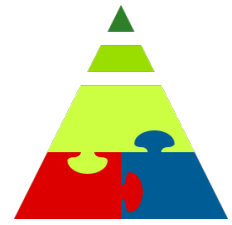
Cloud Computing

Cloud Taking Over Enterprise

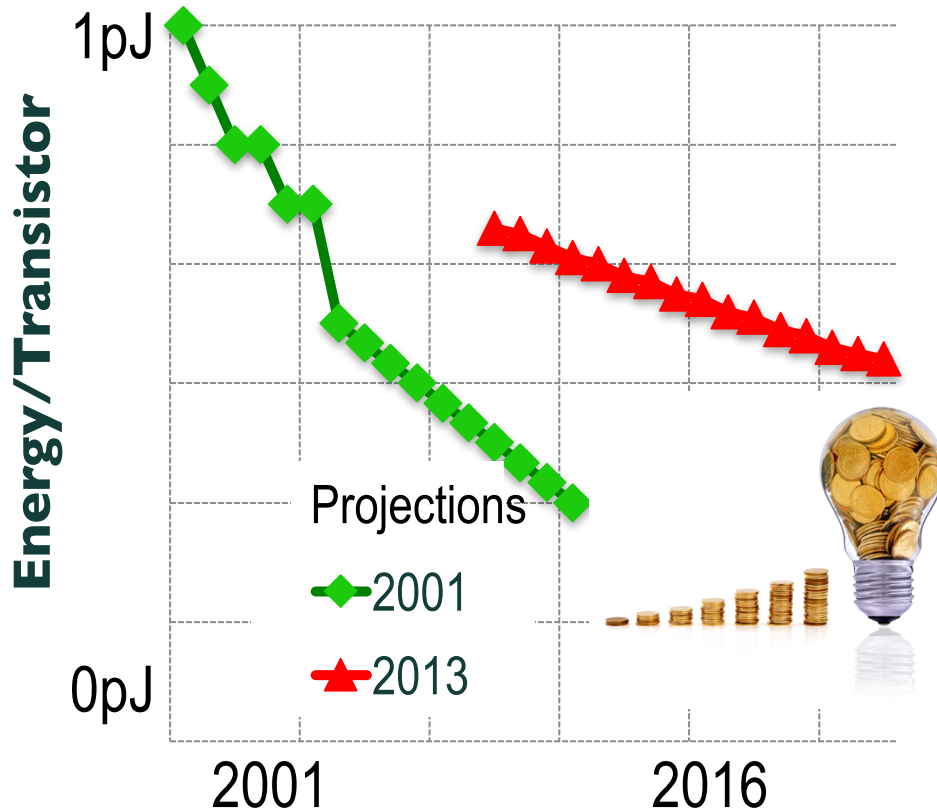


Source: Dell 'Oro 2Q15

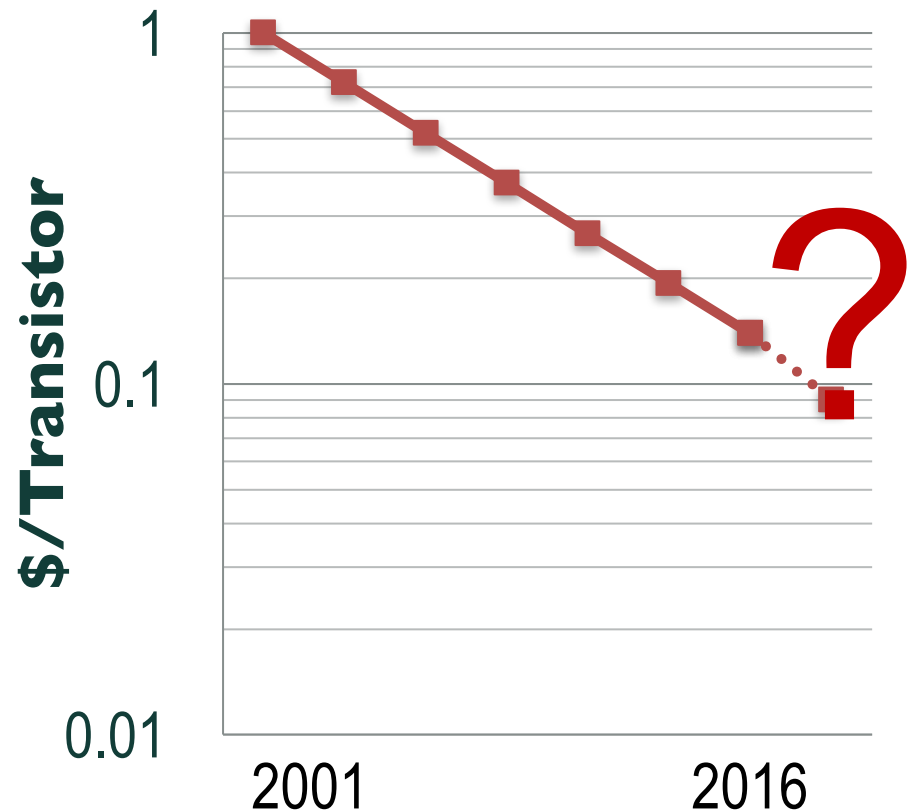
Silicon is running out of steam!



Silicon efficiency is dead
(long live efficient silicon)



Moore's law dying



The Rise of Manycore

With voltages leveling:

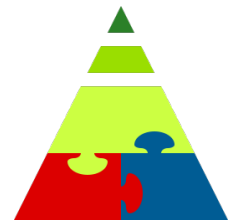
- Parallelism has emerged as the only silver bullet
- Use simpler cores
 - Prius instead of Audi R8
- Restructure software
- Each core → fewer joules/op

Conventional Server
CPU (e.g., Xeon)

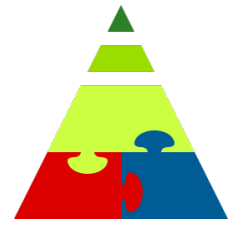


Multicore
Scaling

ARM-based Server
CPU (e.g., ThunderX)



End of Manycore Scaling

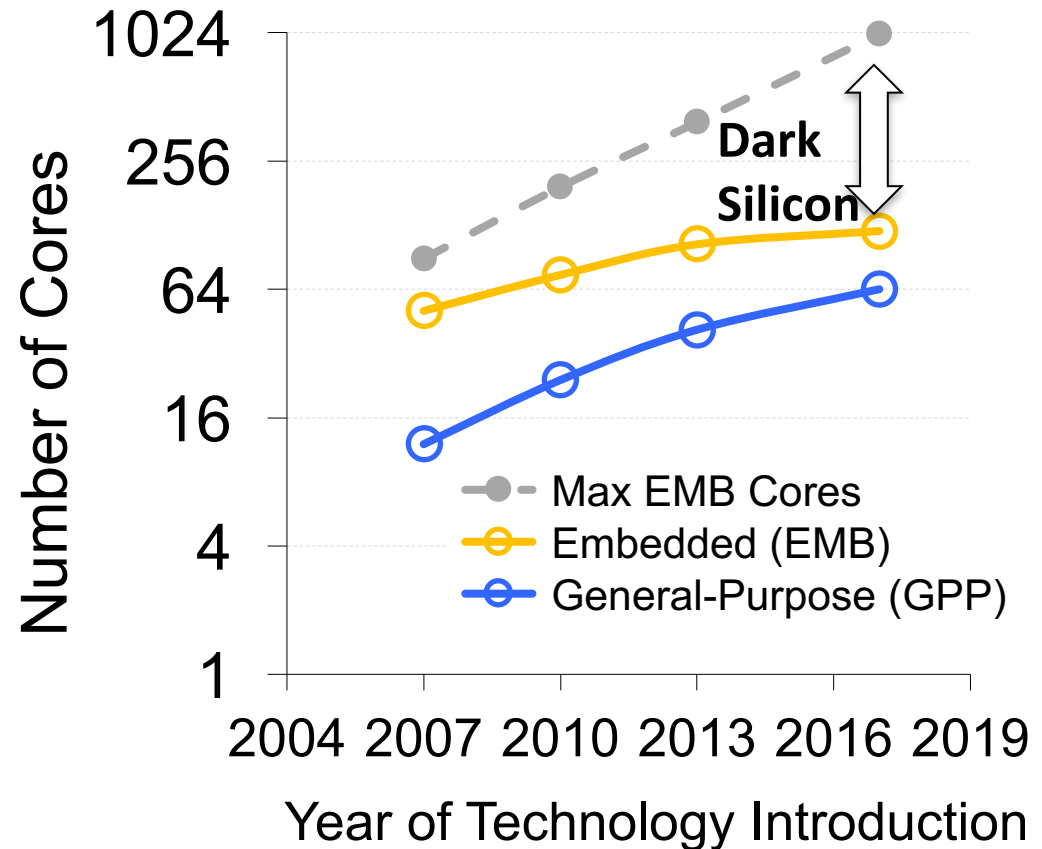


But parallelism can not
offset leveling voltages

Even in servers with
abundant parallelism

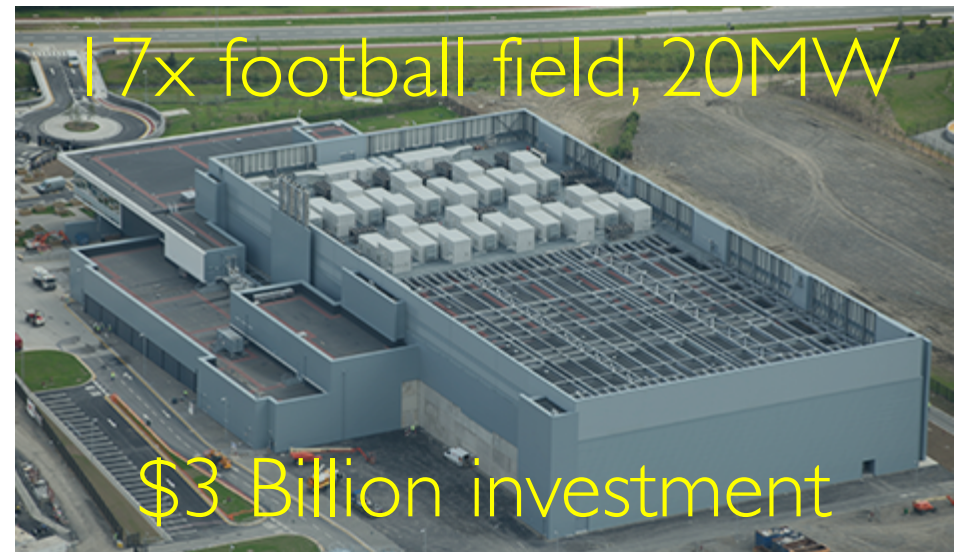
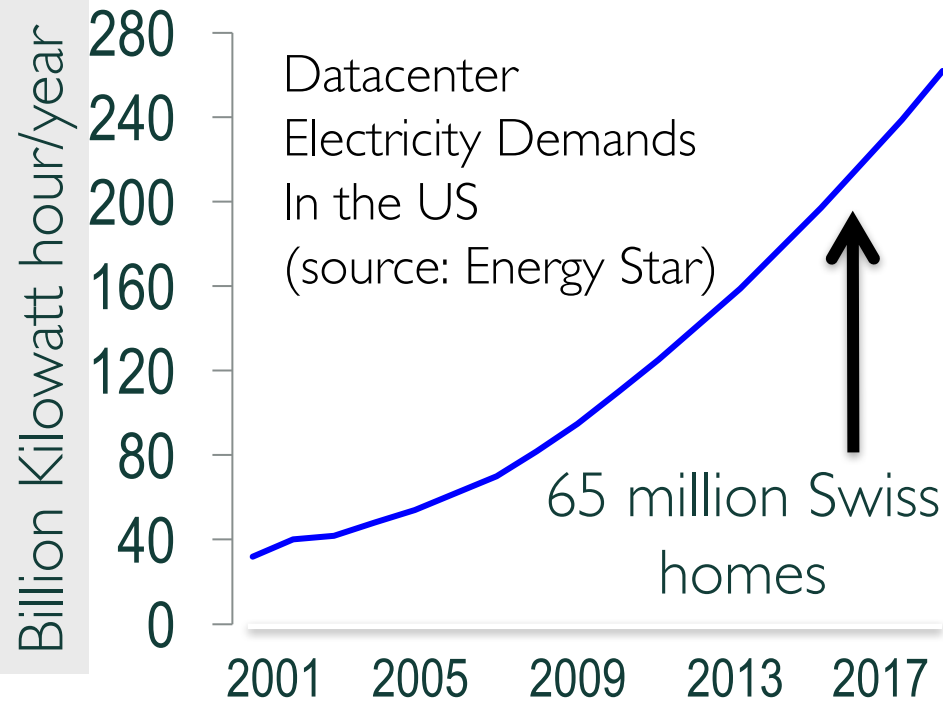
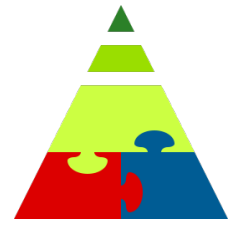
Core complexity has
leveled off too!

Soon, cannot power all chip

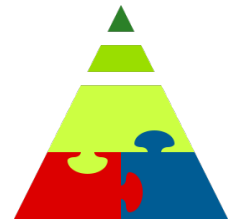


Hardavellas et. al.,
“Toward Dark Silicon in Servers”,
IEEE Micro, 2011

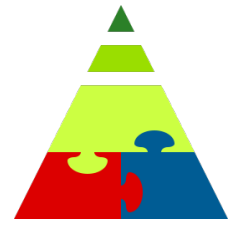
Modern Datacenters are at Physical Limits



- Centralization helps exploit economies of scale
- But, platform scaling is a grand challenge



How do we continue scaling Data Centers?



DIVIDEND:

Distributed Heterogeneous Vertically-Integrated Data Centers

Distributed:

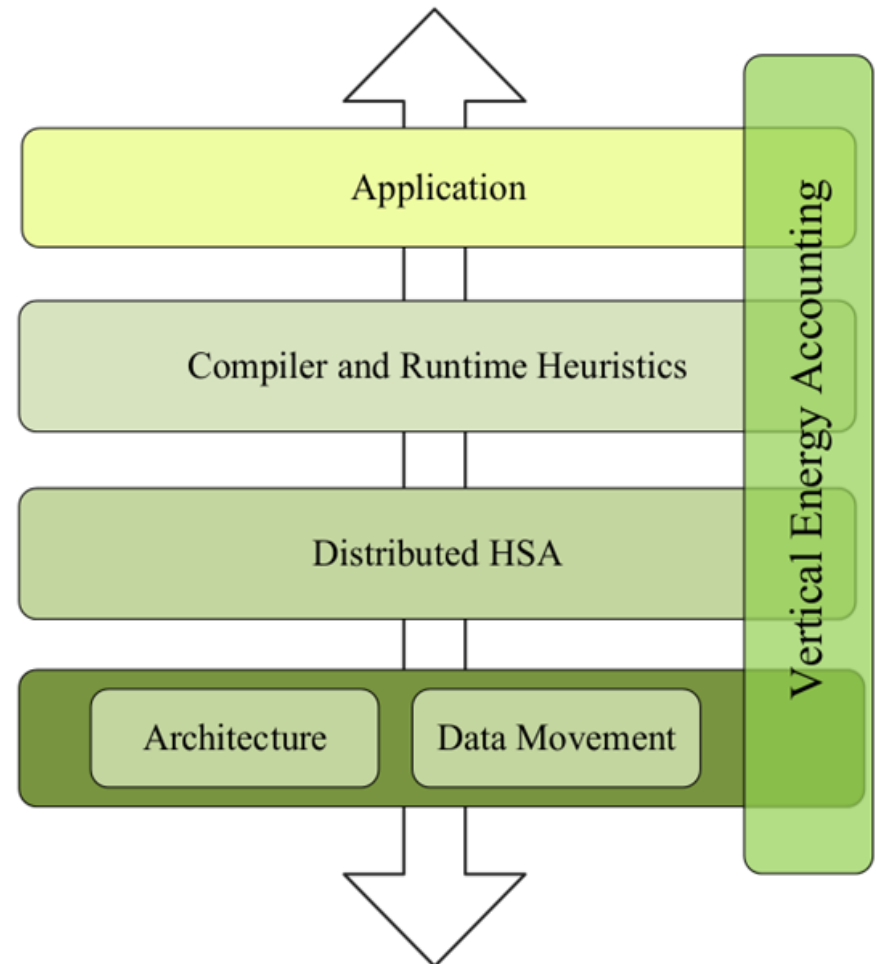
- Scale-out nodes

Heterogeneous:

- Silicon specialization

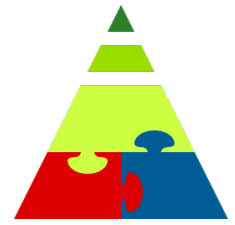
Vertically-integrated:

- Resource monitoring
- Co-design
- Cross-layer optimization

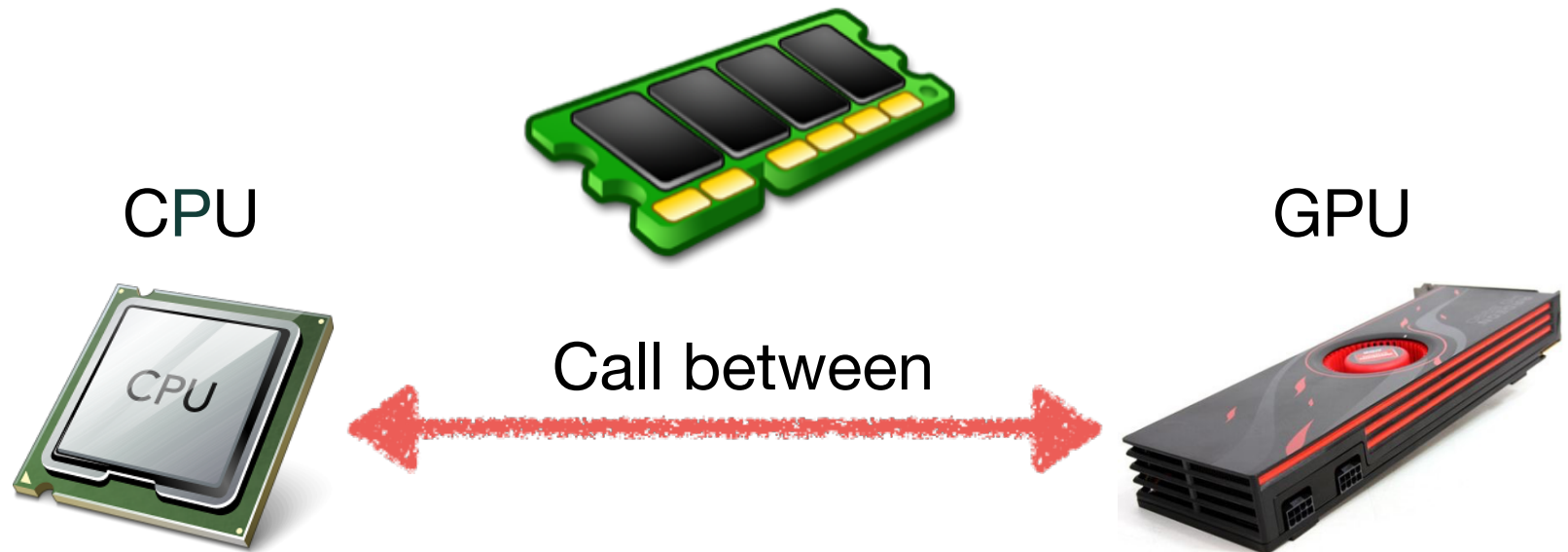


HSA:

Heterogeneous Systems Architecture

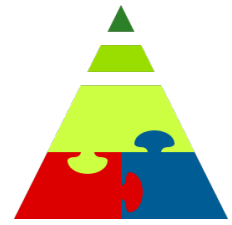


Single memory address space

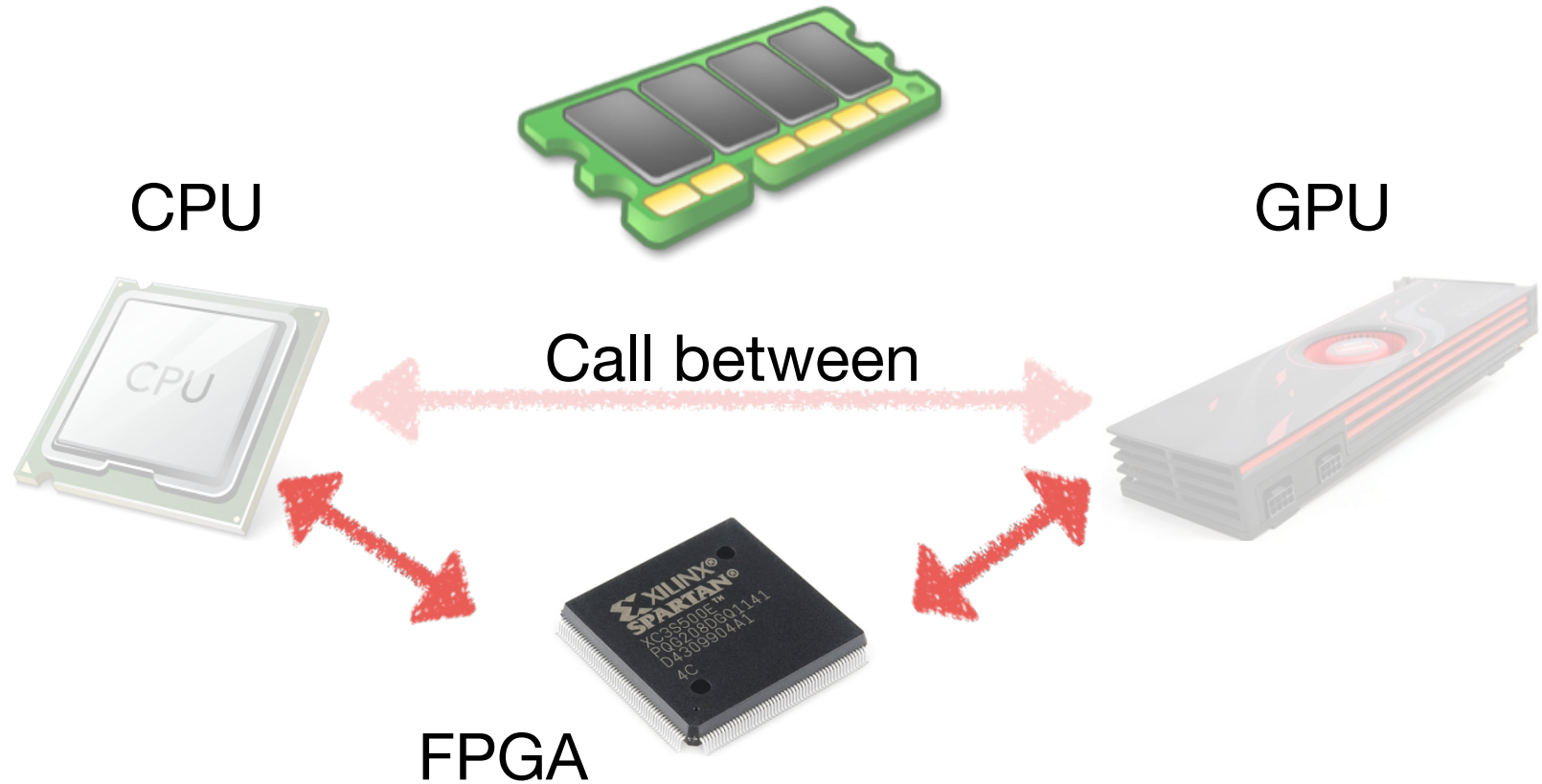


Single programming paradigm

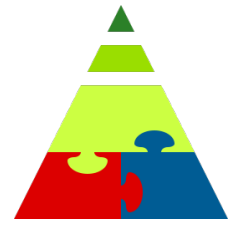
HSA + FPGA



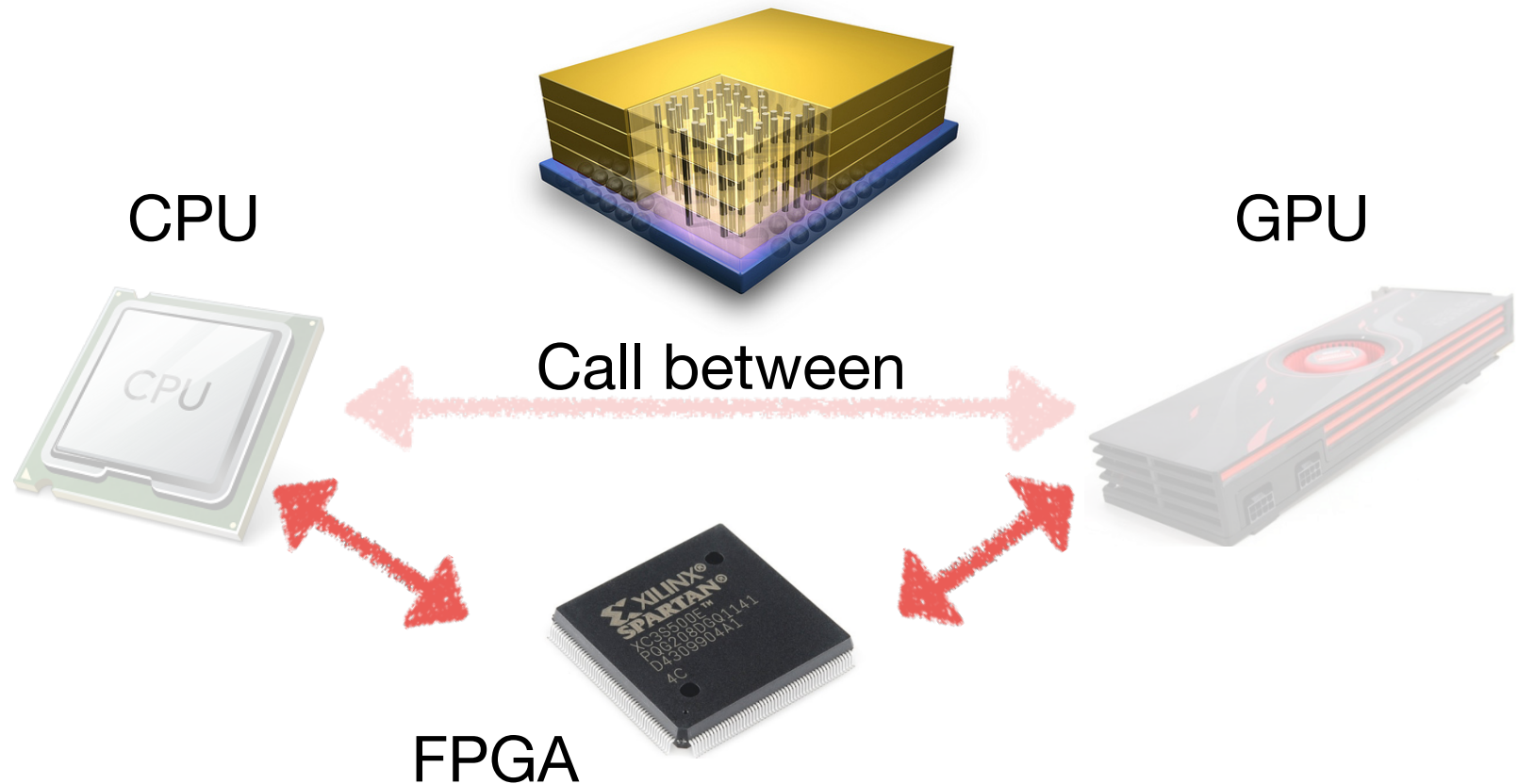
Single memory address space



HSA + FPGA + HMC

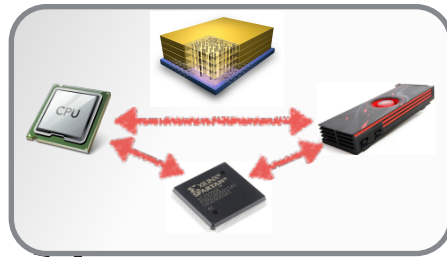
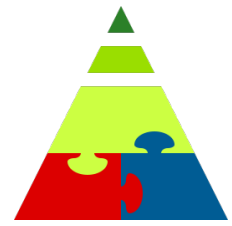


Single address space w/ in-memory accelerators



DHSA

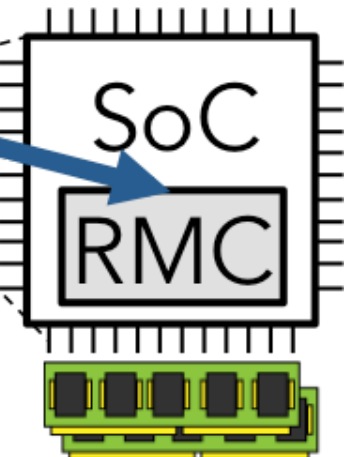
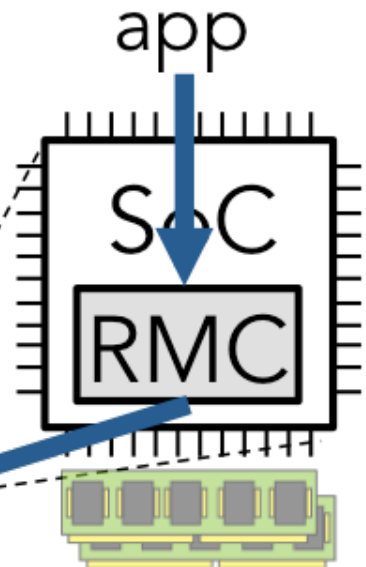
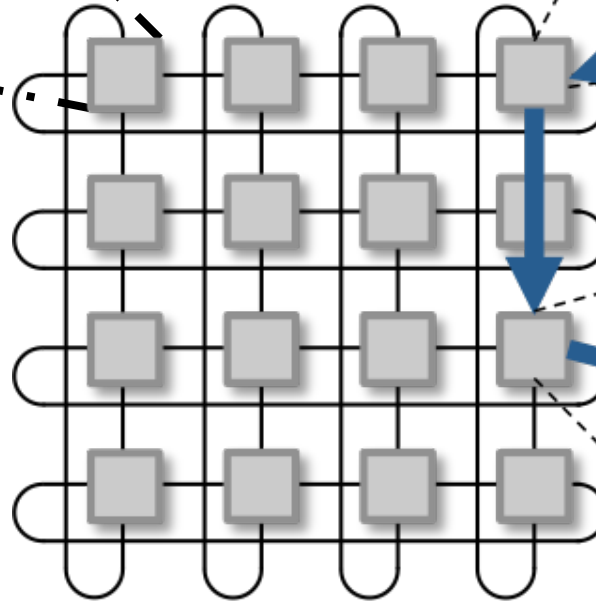
Distributed HSA



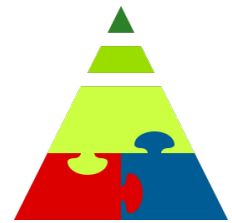
SO-NUMA

Protected remote
access and RPC

Low latency

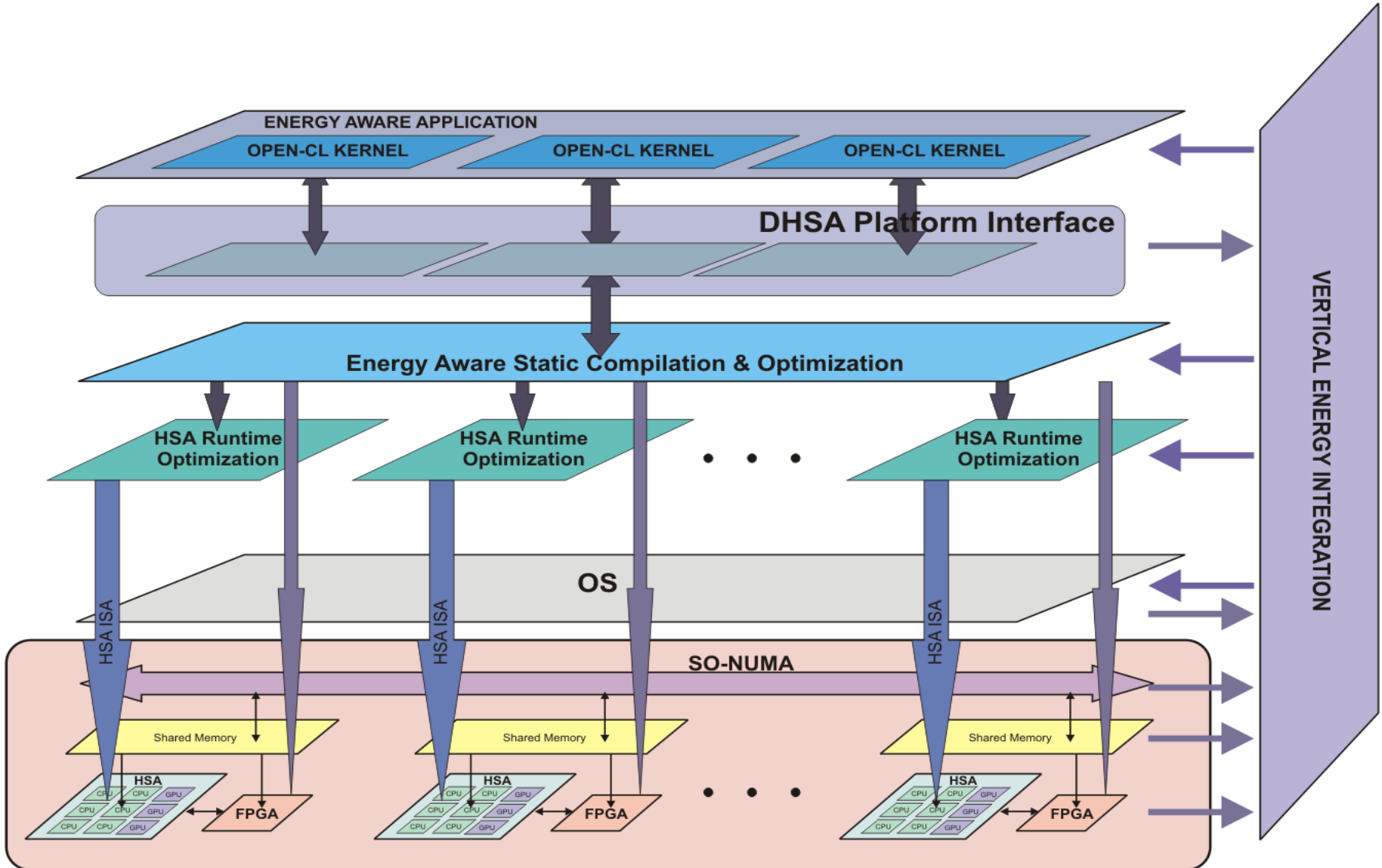
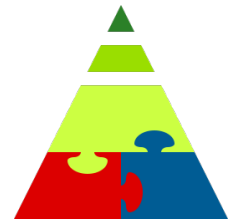


Energy Accounting

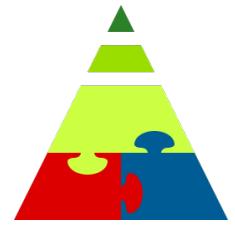


```
QModelIndex start;  
if (currentIndex().isValid())  
    start = currentIndex();  
else  
    start = d->model->index(0, 0, d->root);  
  
bool skipRow = false;  
bool keyboardTimeWasValid = d->keyboardInputTime.isValid();  
qint64 keyboardInputTimeElapsed = d->keyboardInputTime.restart();  
if (search.isEmpty() || !keyboardTimeWasValid  
    || keyboardInputTimeElapsed > QApplication::keyboardInputInterval()) {  
    d->keyboardInput = search;  
    skipRow = currentIndex().isValid(); //if it is not valid we should reali  
} else {  
    d->keyboardInput += search;  
}
```

Project organization



Project Partners



University of Edinburgh

- Hugh Leather

QUB

- Dimitris Nikolopoulos

University of Lancaster

- Zheng Wang

EPFL

- Babak Falsafi

AMD

- Mauricio Breternitz

UPT

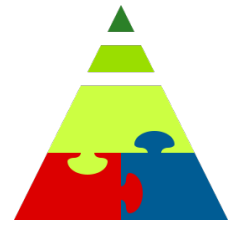
- Alexandru Amaricaei

INRIA

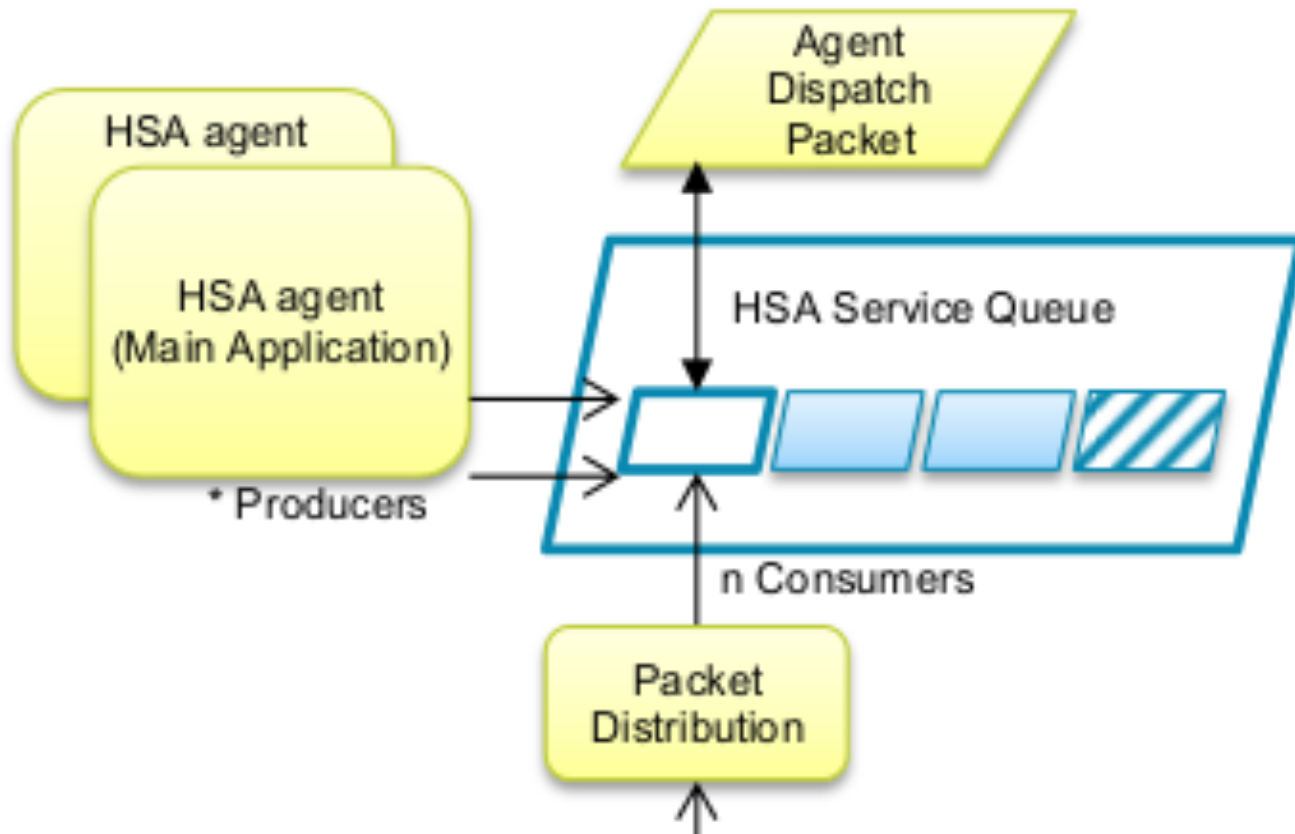
- Albert Cohen

With an army of collaborators at these institutions

DHSA Extension



- Service Queue
- SO-NUMA extension
- HSA-compliant
- Prototype MPI/ethernet



DHSA Energy Accounting

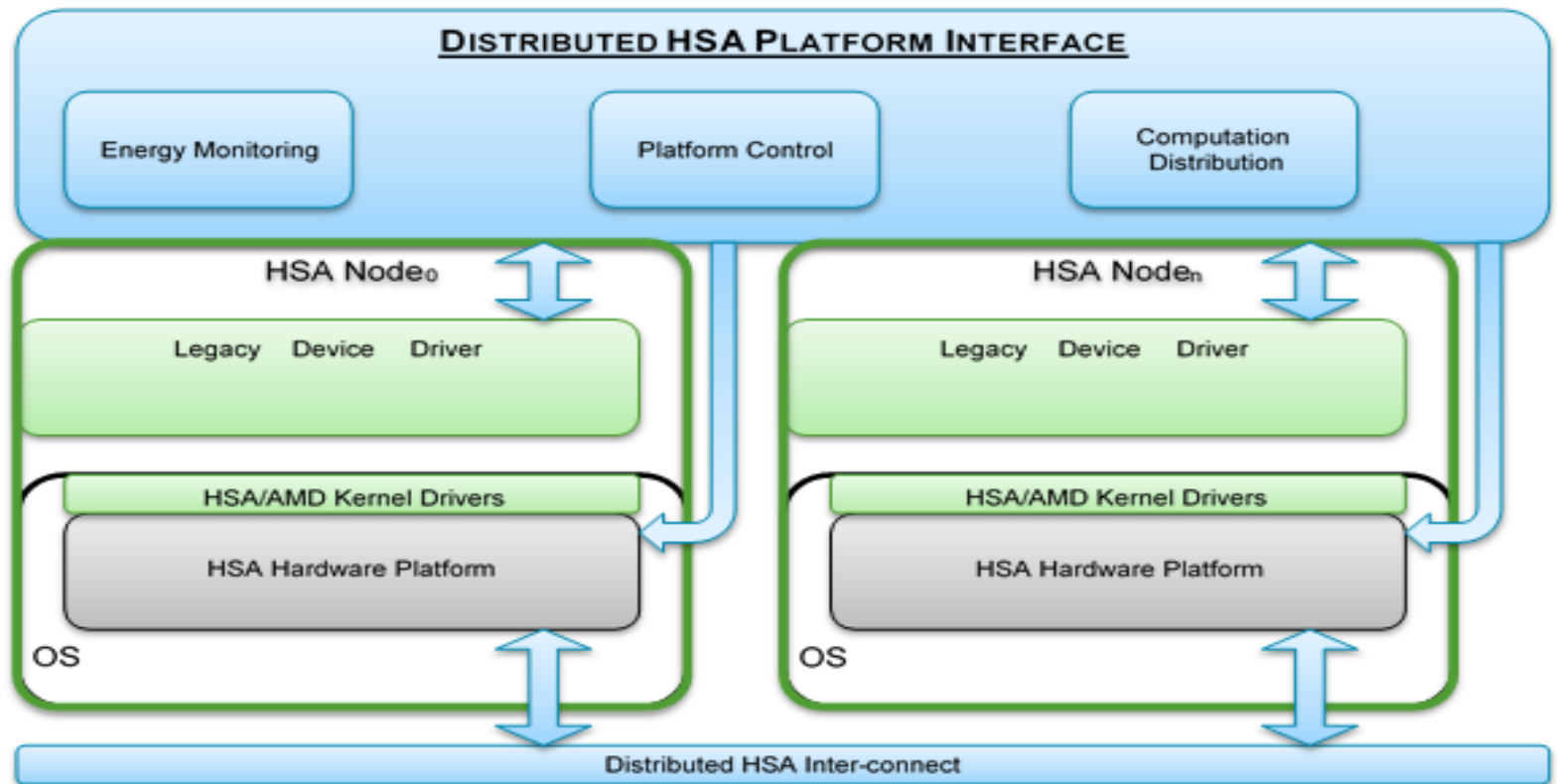


Driver:

- Runtime interface

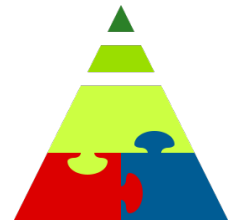
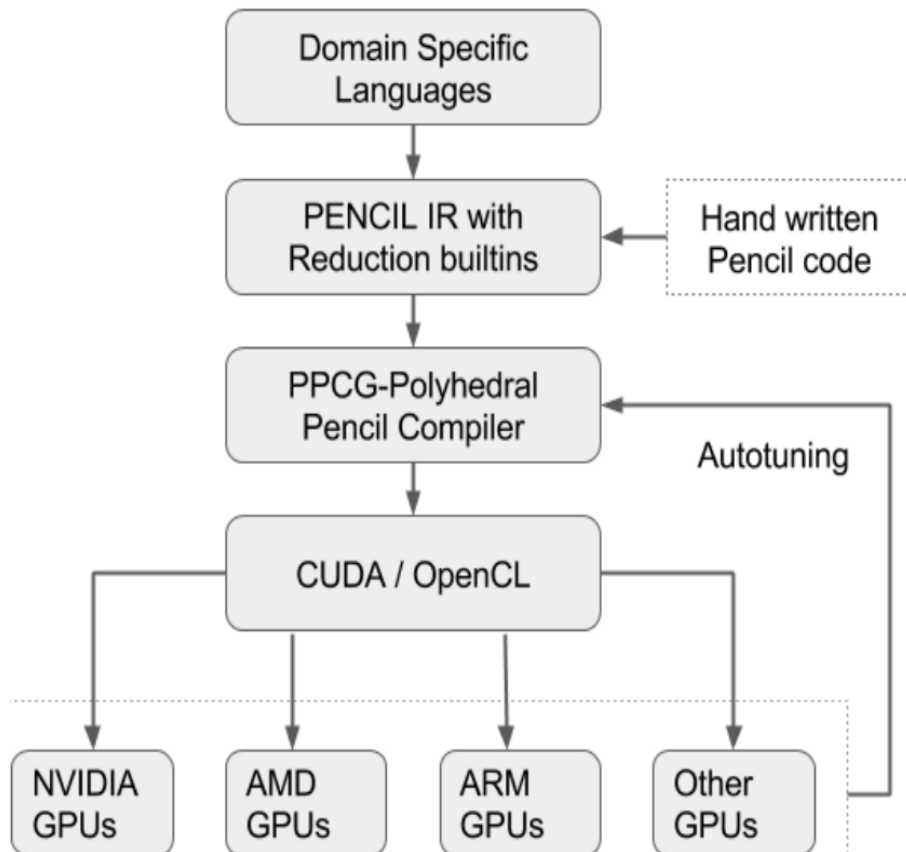
Platform:

- Multiple HSA nodes

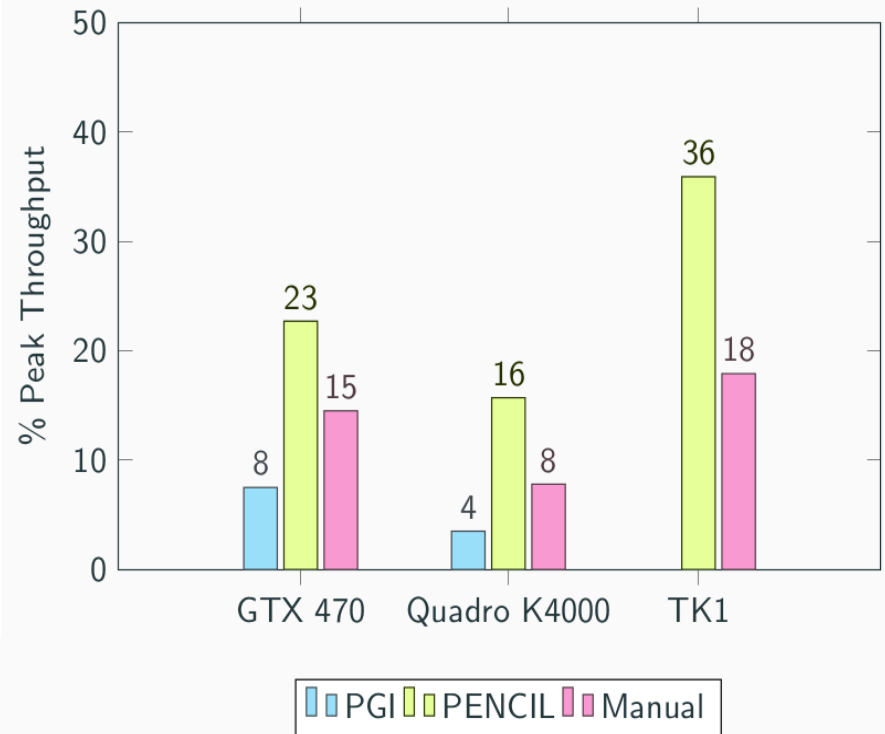


Synthesized Kernels

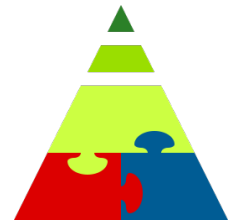
- Accelerated vision apps
- DSL $\rightarrow$ PENCIL $\rightarrow$ OpenCL



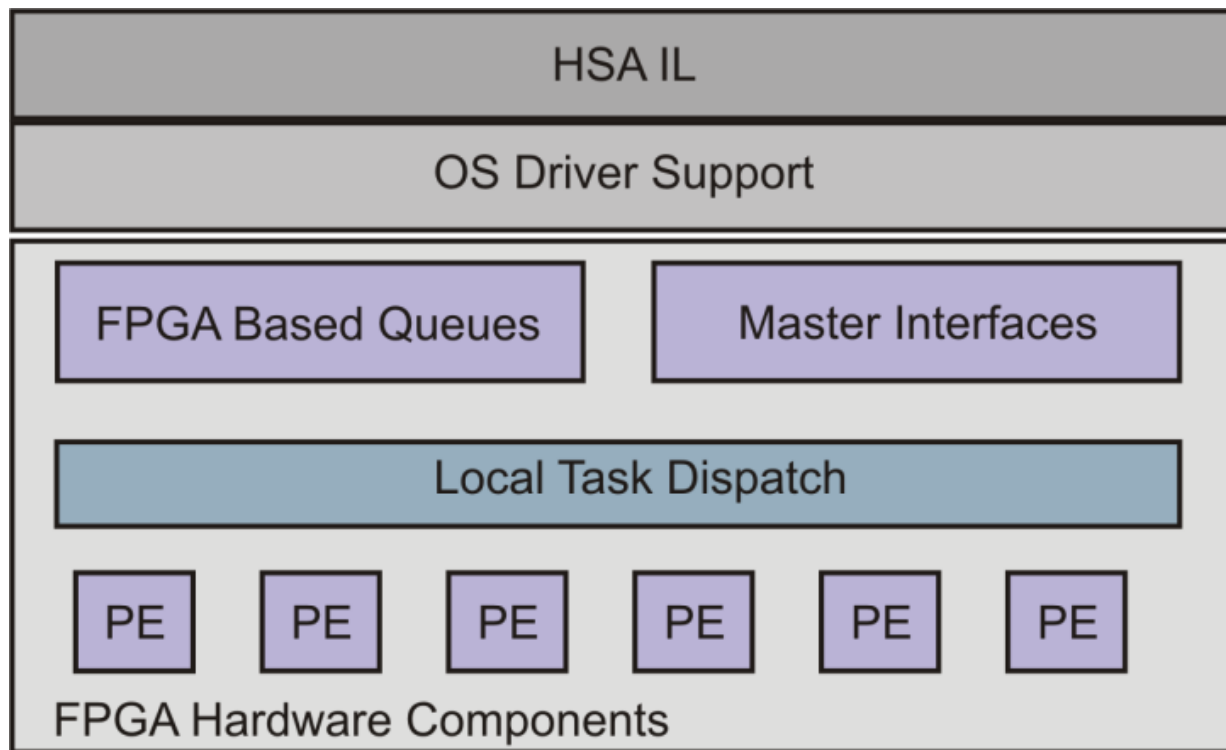
- Example: SLAMbench [Reddy et al., PACT'16]



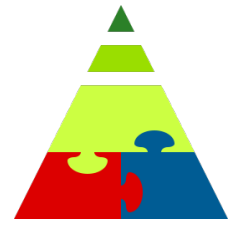
HSA on FPGA



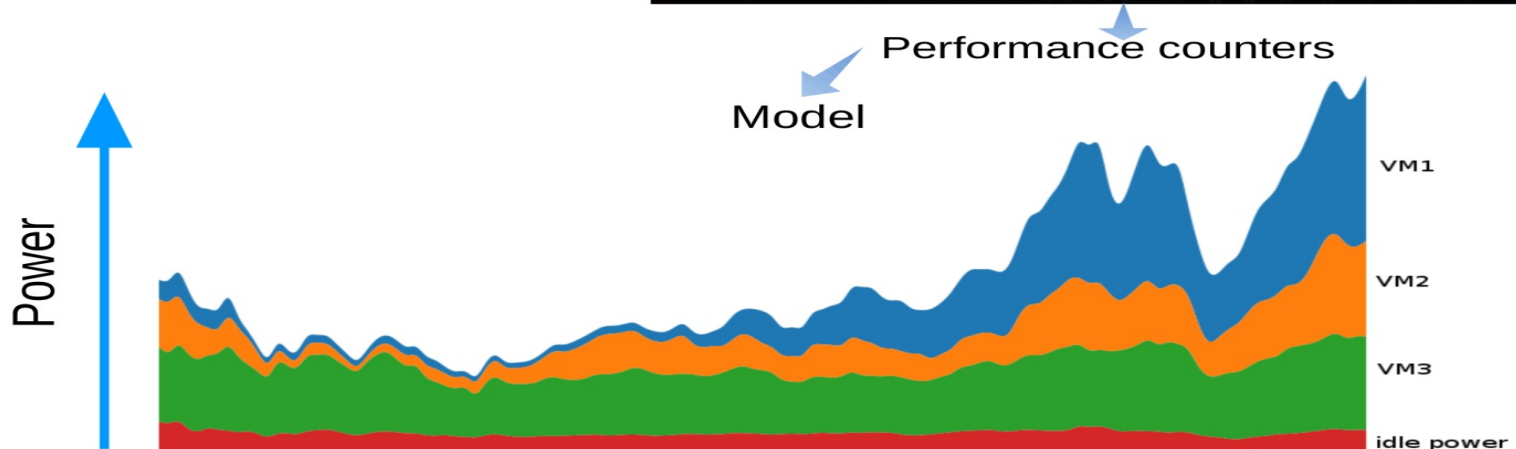
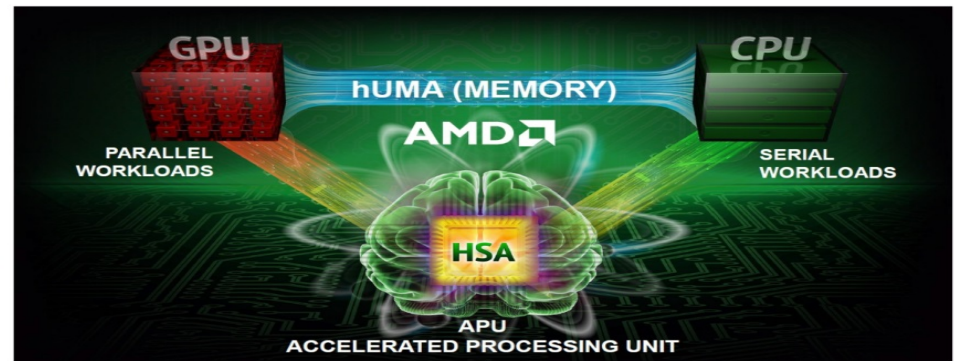
- Prototype Zynq-7020 board
 - 2 ARM Cores, FPGA
 - Connected via an AXI Bus
- Matrix-Vector multiplication
- Barnes-Hut n-body simulation



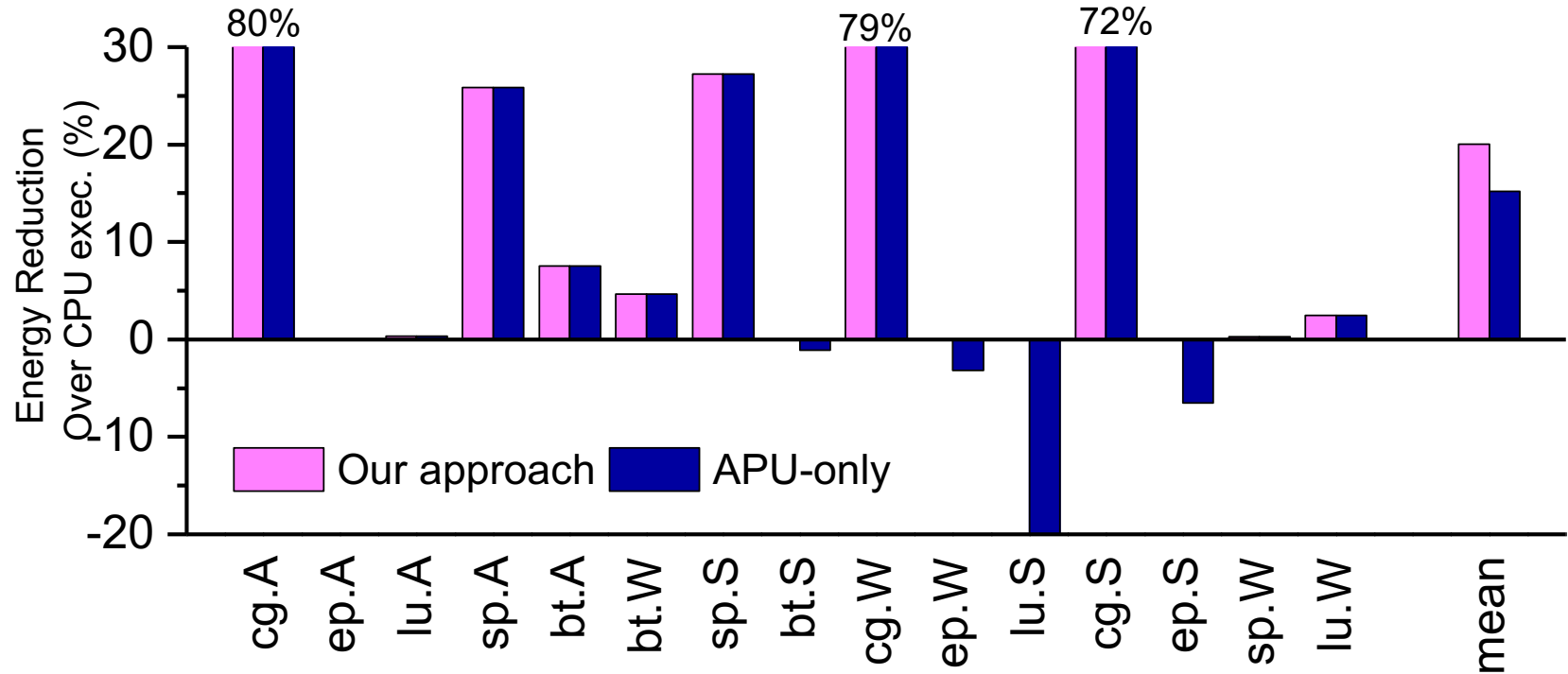
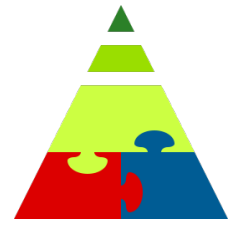
VM Power/Energy Accounting



- VM activity unique → power varies across VM
- Predictive model for VM power @ 1s time within 3 % error
- Energy profiling tool (ALEA) ported to AMD APU

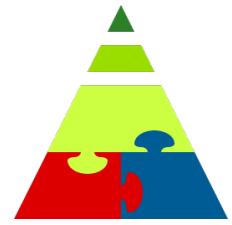


Decision-Tree Based Task Allocation



- 20% reduction on average
- 5x speedup on NVIDIA GTX 580 GPU (not shown)

SABRes: Atomic Remote Object Reads



One-sided reads semantically limited

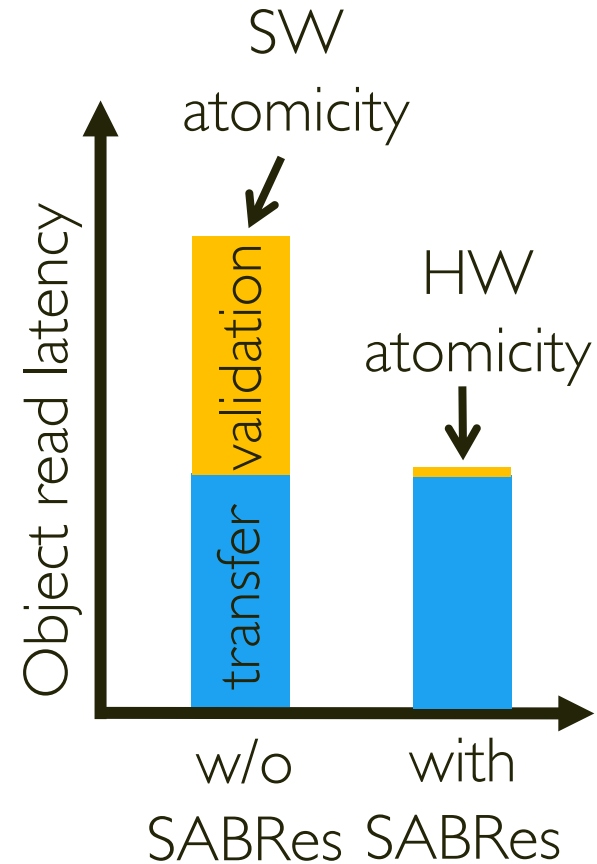
- Atomicity limited to single cache block

Object atomicity through costly SW mechanisms

- Up to 2x higher end-to-end latency

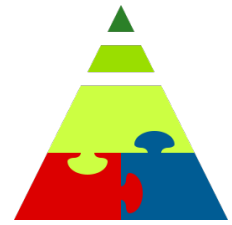
SABRes: HW extension for object atomicity

- Leverage NI's coherent integration
- Remove software overhead

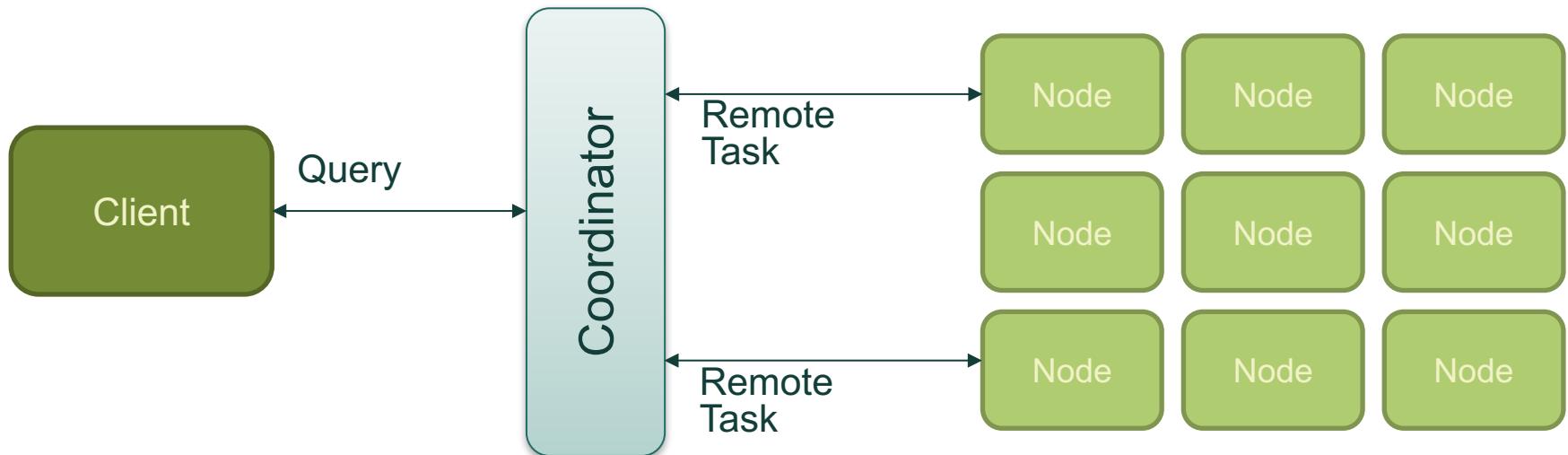


Up to 2x faster distributed object stores

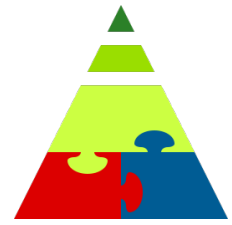
Distributed Analytics



- Dataset types:
 - array-store
 - row-store
 - column-store
- Task: Spark Operators
- Status:
 - ✓ Native/OpenCL for CPU & GPU
 - ✓ DSL for analytics
 - ✓ Port of TPC-H queries
 - ✓ Benchmarking against PostgreSQL



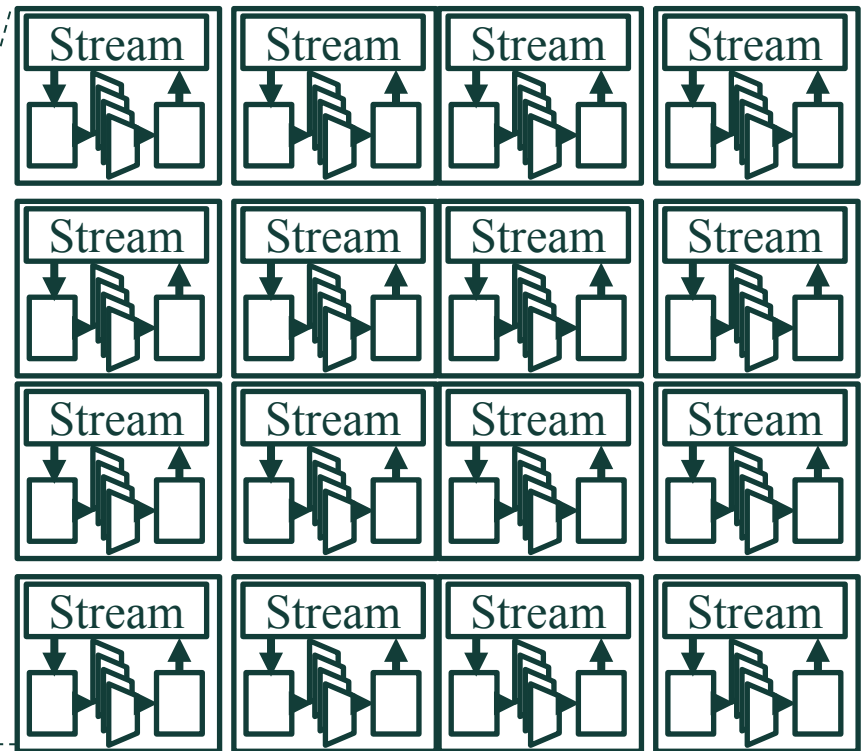
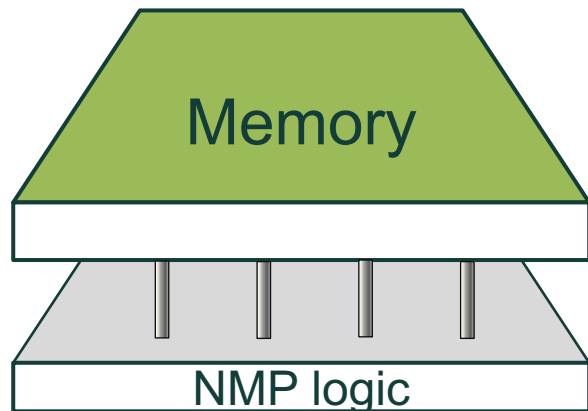
The Mondrian Data Engine



SIMD cores + data streaming

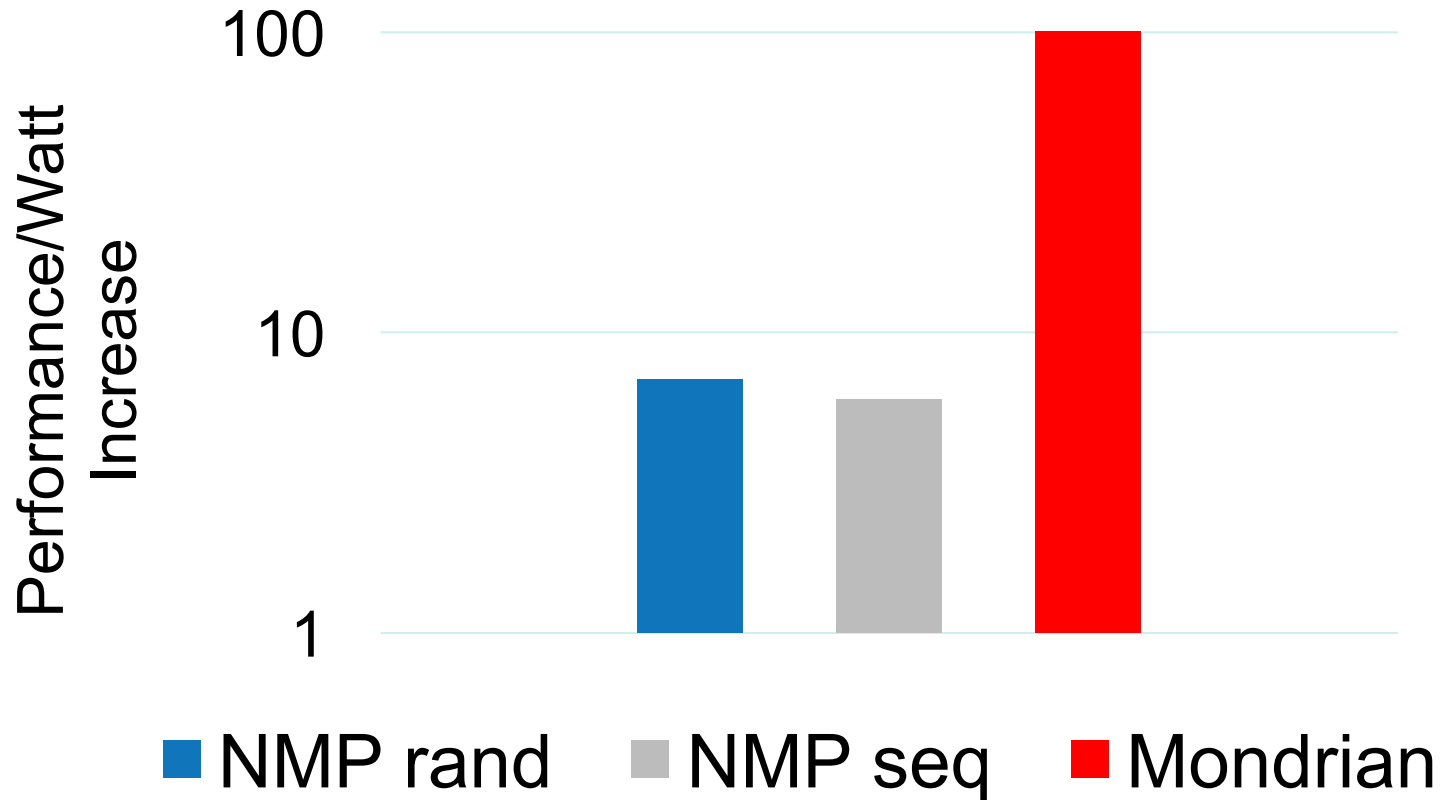
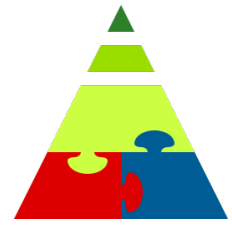
- Streams multiple sequential streams
- 8-wide cores @ 1 GHz
- No caches
- Reaches 8 GB/s per core

Custom ISA



Algorithm/hardware co-design

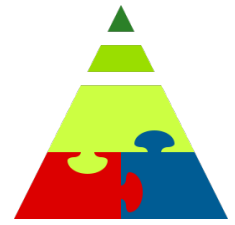
Mondrian Results



- Algorithm alone gets $\sim 10\times$ [ASBD'15]
- Algorithm/hardware co-design gets $100\times$

1. *Power Capping: What Works, What Does Not*, Pavlos Petoumenos, et. al. (ICPADS'15)
2. Towards Collaborative Performance Tuning of Algorithmic Skeletons (HLPGPU'16)
3. Towards Template Based CPU-FPGA Acceleration Framework for Irregular Parallel Applications (DSD'16)
4. An investigation on FPGA based energy profiling of multi-core embedded architectures (ICECS'16)
5. Iterative Compilation on Mobile Devices. (ADAPT'16)
6. Autotuning OpenCL Workgroup Size for Stencil Patterns (ADAPT'16)
7. Power Capping: What Works, What Does Not (ICPADS'15)
8. Performance and Fault Tolerance of Preconditioned Iterative Solvers on Low-Power ARM Architectures (ERPP'15)
9. Configurable FPGA Architecture for Hardware-Software Merge Sorting (MIXDES'16)
10. Minimizing the cost of iterative compilation with active learning (PACT'16)
11. SABRes: Fast atomic remote reads for rack-scale in-memory computing (MICRO'16)
12. DRET: a system for detecting evil-twin attacks in smart homes (SMARTX'16)
13. Exploiting dynamic scheduling for VM-based code obfuscation (TrustCom'16)
14. The Case for RackOut: Scalable Data Serving Using Rack-Scale Systems (SOCC'16)
15. An Analysis of Load Imbalance in Scale-out Data Serving (SIGMETRICS'16)
16. Highly Parametrizable FPGA Architecture for K-Means Clustering (Informatics&Control'17)
17. The Mondrian Data Engine (ISCA'17)
18. Synthesising Benchmarks for predictive Modelling (best paper at CGO'17)
19. End-to-end Deep Learning of Optimization Heuristics (under review PACT'17)
20. ALEA: a fine-grained energy profiling tool (ACM TACO'17)
21. Minimizing the cost of iterative compilation with active learning (CGO'17)
22. Synthesizing benchmarks for predictive modeling (CGO'17)
23. Optimise web browsing on heterogeneous mobile platforms: a machine learning based approach (INFOCOM'17)
24. Real-Time Power Cycling in Video on Demand Data Centres using Online Bayesian Prediction (ICDCS'17)
25. Cracking Android pattern lock in five attempts (NDSS'17)
26. Exploiting wireless received signal strength indicators to detect evil-twin attacks in smart homes (MobiSYS'17)
27. A convolution BiLSTM neural network model for Chinese event extraction (NLPCC'17)
28. Improving first order temporal fact extraction with unreliable data (NLPCC'17)
29. *VM energy accounting for heterogeneous systems*, Hemant Mehta, Lev Mukhanov, Hans Vandierendonck and Dimitrios S. Nikolopoulos, To be submitted.
30. *Extremely fine-grained power profiling*, Lev Mukhanov, Dimitrios S. Nikolopoulos and Bronis R. de Supinski, To be submitted.

Delivered



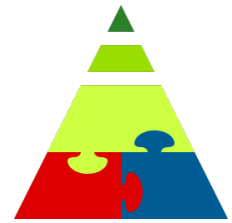
Prototypes

- Open source big data API
- Open source HSAIL auto-tuning
- Open source energy accounting tool
- SO-NUMA NIC logic on Intel HARP
- HSA/OpenCL energy extension

Partnerships

- ARM funded PhD
- INRIA part of HSA Foundation

Summary



- Data centers are at physical limits
- Silicon scaling has slowed down, will stop
- Need server software/hardware
 - Distributed & vertically integrated
 - Novel programming interfaces
 - Energy monitoring/accounting

