DIVIDEND
Heterogeneous Vertically-Integrated Data Centers
Albert Cohen
The future of IT is Data

- Data growth = 100x in ten years [IDC 2012]
  - Population growth = 10% in ten years
- Monetizing data for business, services, health, science
- Big Data is shaping IT & pretty much whatever we do
Science entering 4\textsuperscript{th} 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
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
Modern Datacenters are at Physical Limits

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

Datacenter Electricity Demands
In the US
(source: Energy Star)

<table>
<thead>
<tr>
<th>Year</th>
<th>Datacenter Electricity</th>
<th>65 million Swiss homes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>20 MW</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>30 MW</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>80 MW</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>170 MW</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>280 MW</td>
<td></td>
</tr>
</tbody>
</table>

17x football field 20MW

$3 Billion investment
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

CPU

remote call

GPU

Single programming paradigm
HSA + FPGA

Single memory address space

CPU

remote call

FPGA

GPU
HSA + FPGA + HMC

Single address space w/ in-memory accelerators

CPU

remote call

FPGA

GPU
DHSA
Distributed HSA

Protected remote access and RPC
Low latency
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 Amaricai
INRIA
  ▪ Albert Cohen

With many 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 ➔ PENCIL ➔ 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

![HSA on FPGA Diagram]
VM Power/Energy Accounting

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

- 20% reduction on average
- 5x speedup over 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 $\sim 10x$ [ASBD’15]
- Algorithm/hardware co-design gets $100x$
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)
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)
15. An Analysis of Load Imbalance in Scale-out Data Serving (SIGMETRICS’16)
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 (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)
30. **Extremely fine-grained power profiling**, Lev Mukhanov, Dimitrios S. Nikolopoulos and Bronis R. de Supinski.
Delivered

Prototypes
- Open source API for accelerated analytics
- Open source HSAIL auto-tuning
- Open source energy accounting tool
- SO-NUMA NIC logic on Intel HARP
- HSA/OpenCL energy extension
- DHSA proposal and proposal and evaluation

Partnerships
- EcoCloud consortium
- ARM funded PhDs
- INRIA part of HSA Foundation and with Facebook on deep learning acceleration
Summary

▪ Data centers are at physical limits
▪ Silicon scaling has slowed down, will stop
▪ DIVIDEND’s codesigned software/hardware
  ▪ Distributed & vertically integrated
  ▪ Novel programming interfaces
  → Transparent acceleration (GPU, APU, FPGA)
  → Energy monitoring/accounting
▪ Follow-up industry transfer activities