The Challenges and Opportunities of Micro-Servers in the HPC Ecosystem

Dimitrios S. Nikolopoulos

School of Electronics, Electrical Engineering and Computer Science
Queen's University of Belfast

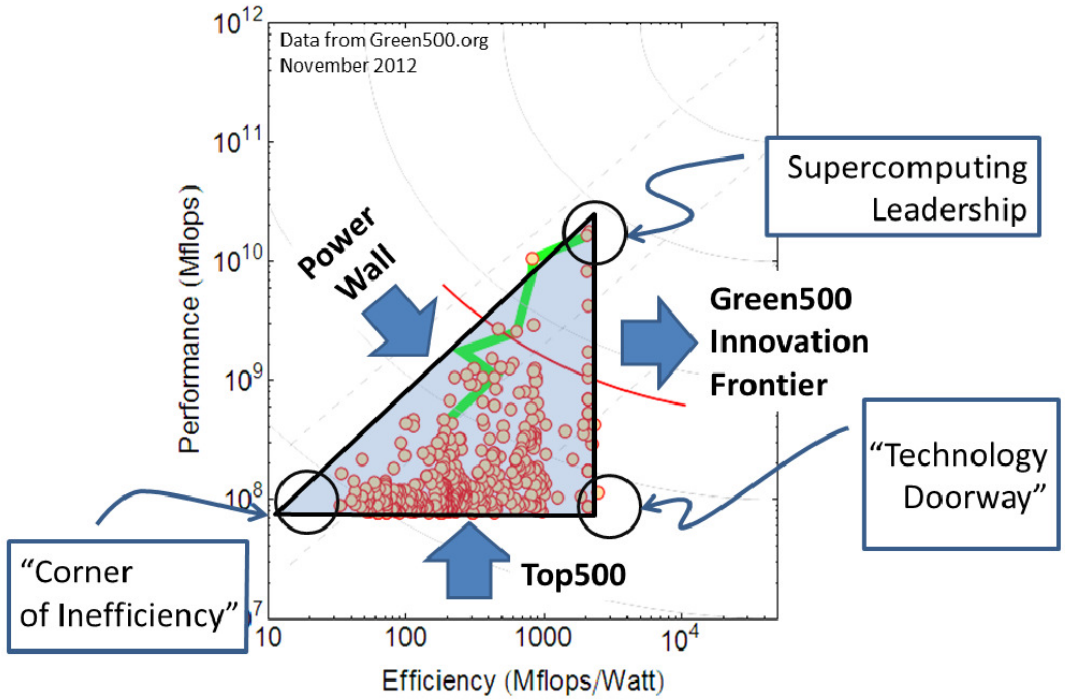
September 4, 2014

Outline

1. HPC and the low-power processor ecosystem
2. The NanoStreams proposition
3. Financial real-time analytics
4. In-memory column stores
5. Conclusions

What we know

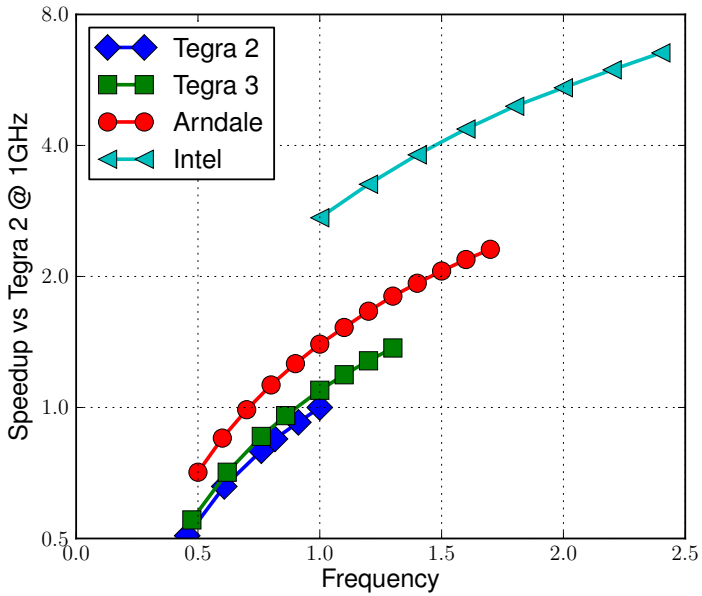
Technology alone can not bridge the gap[1]

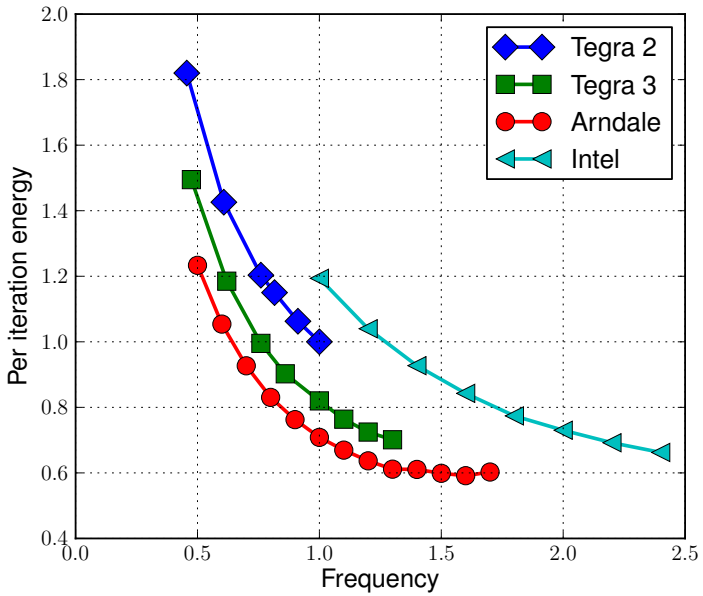


[1] B. Subramaniam, W. Saunders, T. Scogland and W. Feng, Trends in Energy-Efficient Computing: A Perspective from the Green500, International Green Computing Conference (IGCC), 2013, Arlington, Virginia, USA.

HPC and ARM

Single-core ARM[2]

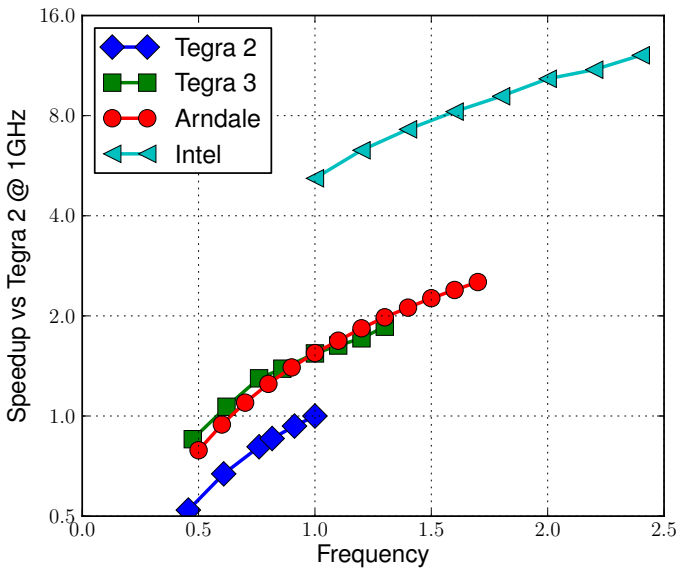




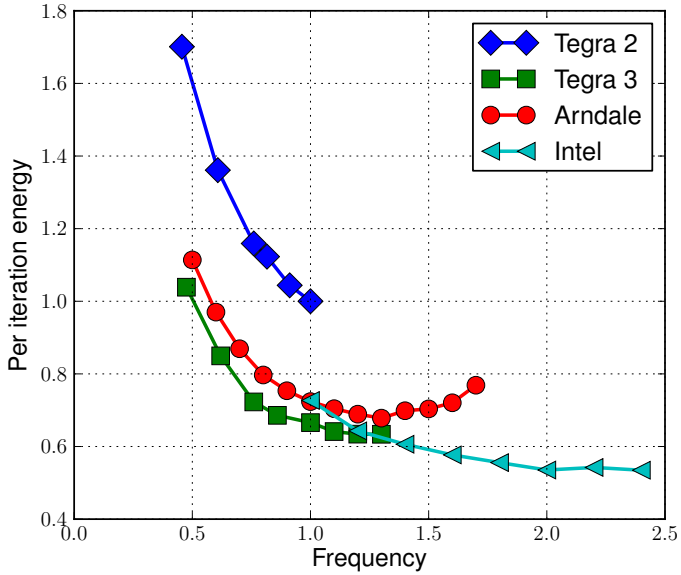
[2] Source: Nikola Rajovic, Paul M. Carpenter, Isaac Gelado, Nikola Puzovic, Alex Ramirez, and Mateo Valero. 2013. Supercomputing with commodity CPUs: are mobile SoCs ready for HPC?. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '13). ACM, New York, NY, USA, , Article 40.

HPC and ARM

...and the performance shortfall



(a) Performance



(b) Energy

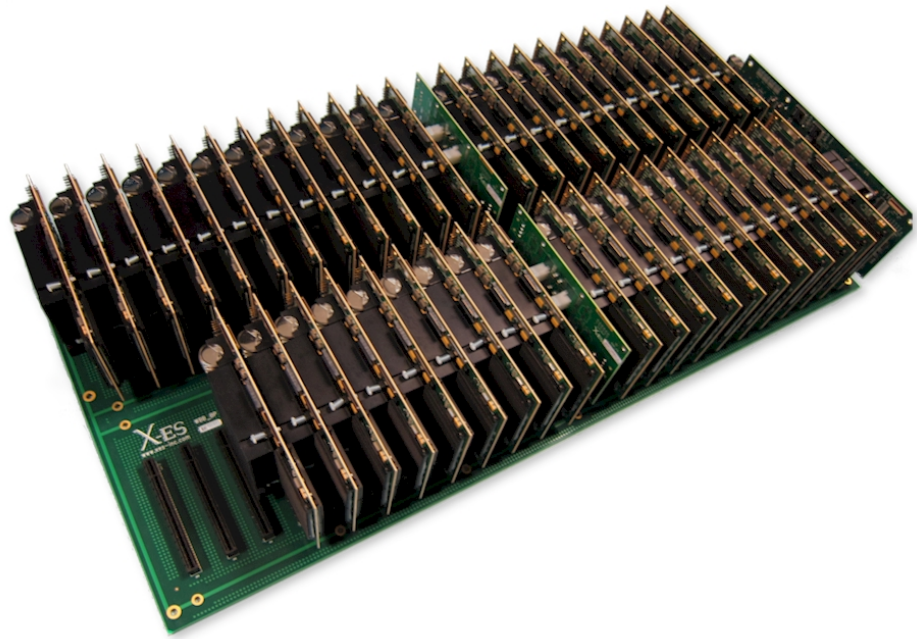
3

[3] Source: Nikola Rajovic, Paul M. Carpenter, Isaac Gelado, Nikola Puzovic, Alex Ramirez, and Mateo Valero. 2013. Supercomputing with commodity CPUs: are mobile SoCs ready for HPC?. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '13). ACM, New York, NY, USA, , Article 40 , 12 pages.

Queen's University Belfast

Microserver concept

- Lightweight and scale-out oriented
  - 1U fits 24–48 cards
- Targeting datacenters, in particular web services
  - no FP, but latency-sensitive
- Shared fan and power supply
- Wide range of processor choices within low power envelopes
- Favoring commodity memory & interconnects (Ethernet vs. IB, LPDDR vs. DDR)



Outline

1 HPC and the low-power processor ecosystem

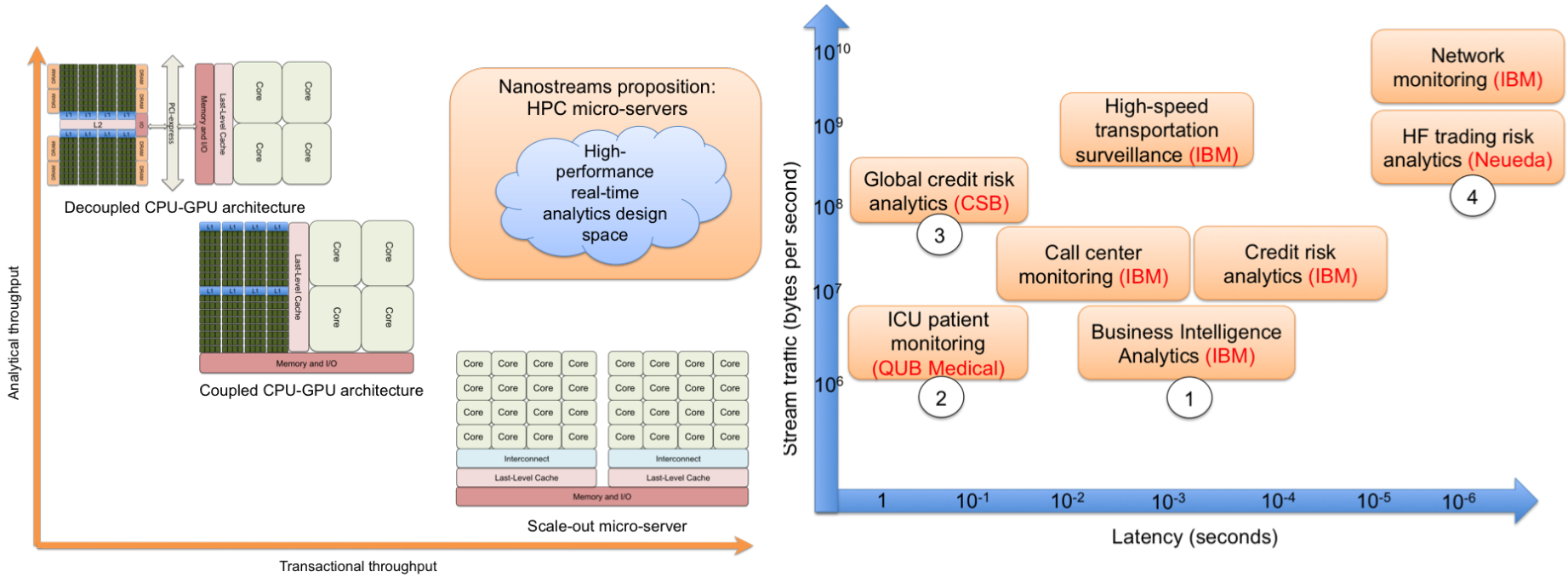
2 The NanoStreams proposition

3 Financial real-time analytics

4 In-memory column stores

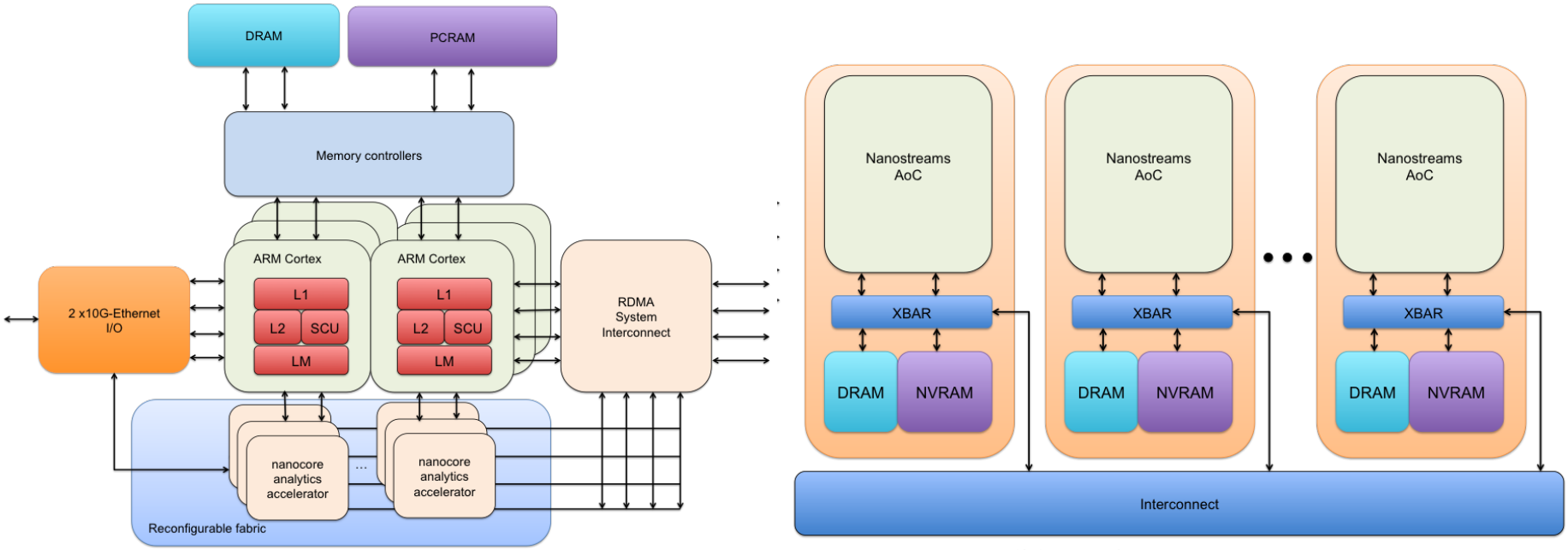
5 Conclusions

Gap in the server landscape[4]



[4] http://www.nanostreams.eu

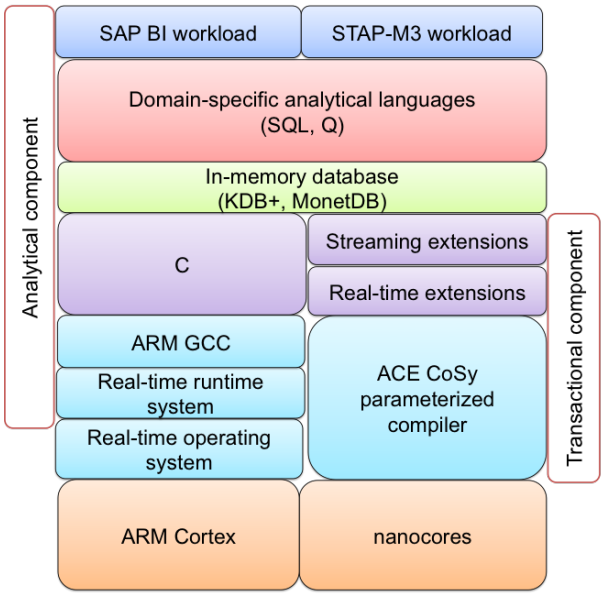
NanoStreams AoC block

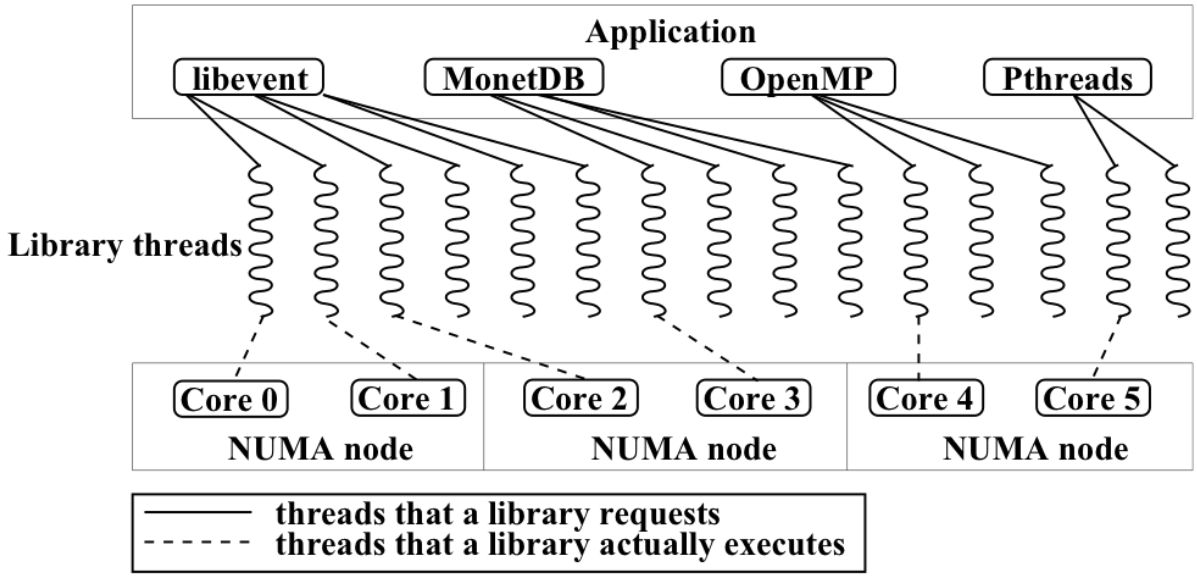


- AoC host on Calxeda boards (A9 cores, 10 GigE)
  - Odroid boards explored as alternative: (A15 cores, GigE)
- AoC accelerator on Xilinx Zynq boards

NanoStreams software stack

Taming oversubscription and latency





- Space and time isolation of parallel components
- RDMA over raw Ethernet, user-level
- Soft real-time scheduling guarantees
- Locality exploitation both horizontally and vertically

Queen's University Belfast

Outline

1 HPC and the low-power processor ecosystem

2 The NanoStreams proposition

3 Financial real-time analytics

4 In-memory column stores

5 Conclusions

Option pricing

- Datacenters co-located with trading venues
- No flexibility in moving the datacenter "where electricity is cheap"
- No flexibility in running the datacenter "when electricity is cheap"
- Not particularly compute- or data-intensive, low-latency workloads
  - Monte Carlo simulations, Black Scholes, Binomial Pricing
  - Instance runs in ms or μs
  - Heavily traded symbols trigger Koptions/session

$$\text{Price} = (-1)^p \left(SN((-1)^p d_1) - Pe^{-rT} N((-1)^p d_2) \right) \tag{1}$$

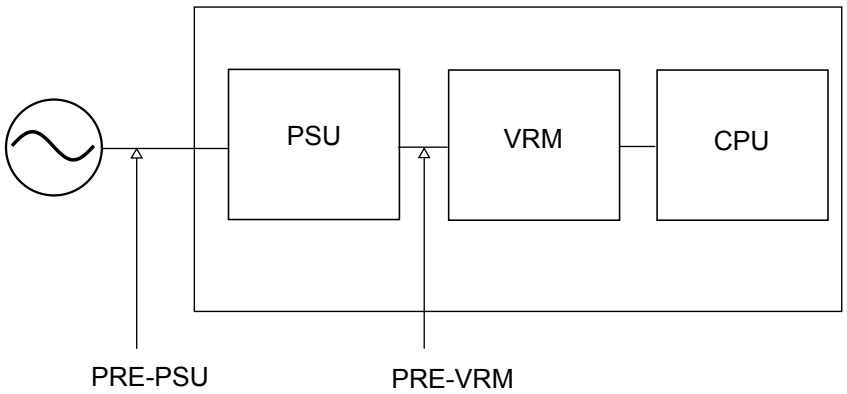
$$\text{Price} = \frac{e^{-rT}}{N} \sum_{i=1}^{N} \max \left(0, S - Pe^{(r-\frac{\sigma^2}{2})T + \sigma\sqrt{T}x_i} \right) \tag{2}$$

$$u = e^{\sigma\sqrt{T}} \quad \text{and} \quad d = \frac{1}{u} \tag{3}$$

Energy-efficiency metrics and measurement approaches

Real-time, latency-sensitive workloads[5]

- **Joules/option**: Provider-side, sustained throughout trading day, reduction translates to less TCO
- **Time/option**: User-side, end-to-end latency.
- **QoS:** Calculating option before new price arrives; unknown deadline.



[5] Charles Gillan, Dimitrios S. Nikolopoulos, Giorgis Georgakoudis, Richard Faloon, George Tzenakis and Ivor Spence: On the Viability of Micro-Servers for Financial Analytics, *TR:HPDC-RC:2014:08:29*.

Scale-out pays off?

Dell (Intel Sandybridge) vs. Boston Viridis (ARM Cortex) servers

Replayed, real, trading day market feed with 617 option pricing instances on Facebook stock

Table : Power profiles for standalone kernel kernels

Kernel and Platform	N	PRE-VRM $\bar{P}$(W)	Time (s)	J/opt
MC Intel	0.5M	25.8	8.6	0.36
	2.0M	26.0	34.0	1.37
MC Viridis(1)	0.5M	6.8	41.2	0.45
	2.0M	7.4	163.7	1.96
MC Viridis(16)	0.5M	108.8	2.9	0.51
	2.0M	118.4	10.1	1.94
BT Intel	4000	24.5	8.6	0.34
	7000	24.9	32.8	1.86
BT Viridis(1)	4000	5.0	42.0	0.35
	7000	5.2	132.0	1.07
BT Viridis(16)	4000	88.0	2.8	0.40
	7000	97.6	8.0	1.27

Session-wide energy efficiency

Table : J/opt for execution of the standalone kernels using the PRE-PSU power measurement

	N	Intel		Viridis(1)		Viridis(16)	
		$\bar{P}$(W)	J/opt	$\bar{P}$(W)	J/opt	$\bar{P}$(W)	J/opt
MC	0.5M	109.1	1.52	136.3	9.1	238.3	1.12
	1.0M	112	3.16	135.5	18.1	244.6	2.22
	2.0M	114.1	6.29	134.9	35.8	245.0	4.01
BT	4000	109.8	1.53	135.2	9.2	245.4	1.11
	5000	111.7	2.68	135.4	14.4	244.7	1.67
	7000	112.1	5.96	135.1	28.9	245.3	3.18

Queen's University Belfast

How QoS changes the overall picture

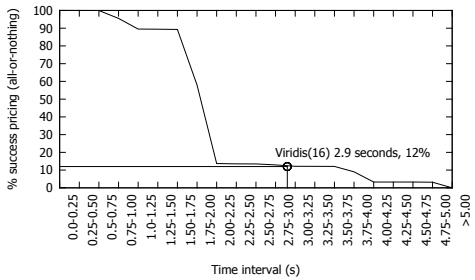


Table : QoS metric and TCO in various setups

MC 1M	QoS	# Options priced	PRE-PSU $\bar{P}$(W)	TCO KWh
Intel(1)	13.2%	827593	112.0	0.73
Viridis(1)	2.6%	162873	135.5	0.88
Viridis(2)	5.2%	325048	141.9	0.92
Viridis(4)	10.4%	649402	158.0	1.03
Viridis(8)	20.8%	1305408	187.5	1.22
Viridis(16)	41.5%	2600416	244.6	1.59
***Intel(2)**	26.4%	1655186	224.0	1.46
***Intel(3)**	39.6%	2482779	336.0	2.18

Outline

1 HPC and the low-power processor ecosystem

2 The NanoStreams proposition

3 Financial real-time analytics

4 In-memory column stores

5 Conclusions

Modeling the Energy of NVRAM[6]

- NVRAM is viable DRAM alternative with DRAM failing to scale beyond 22 nm
- Various options: PCM, STT-RAM, RRAM

$$T(L) = \frac{N}{\phi}(CPI_0 + ML) \tag{4}$$

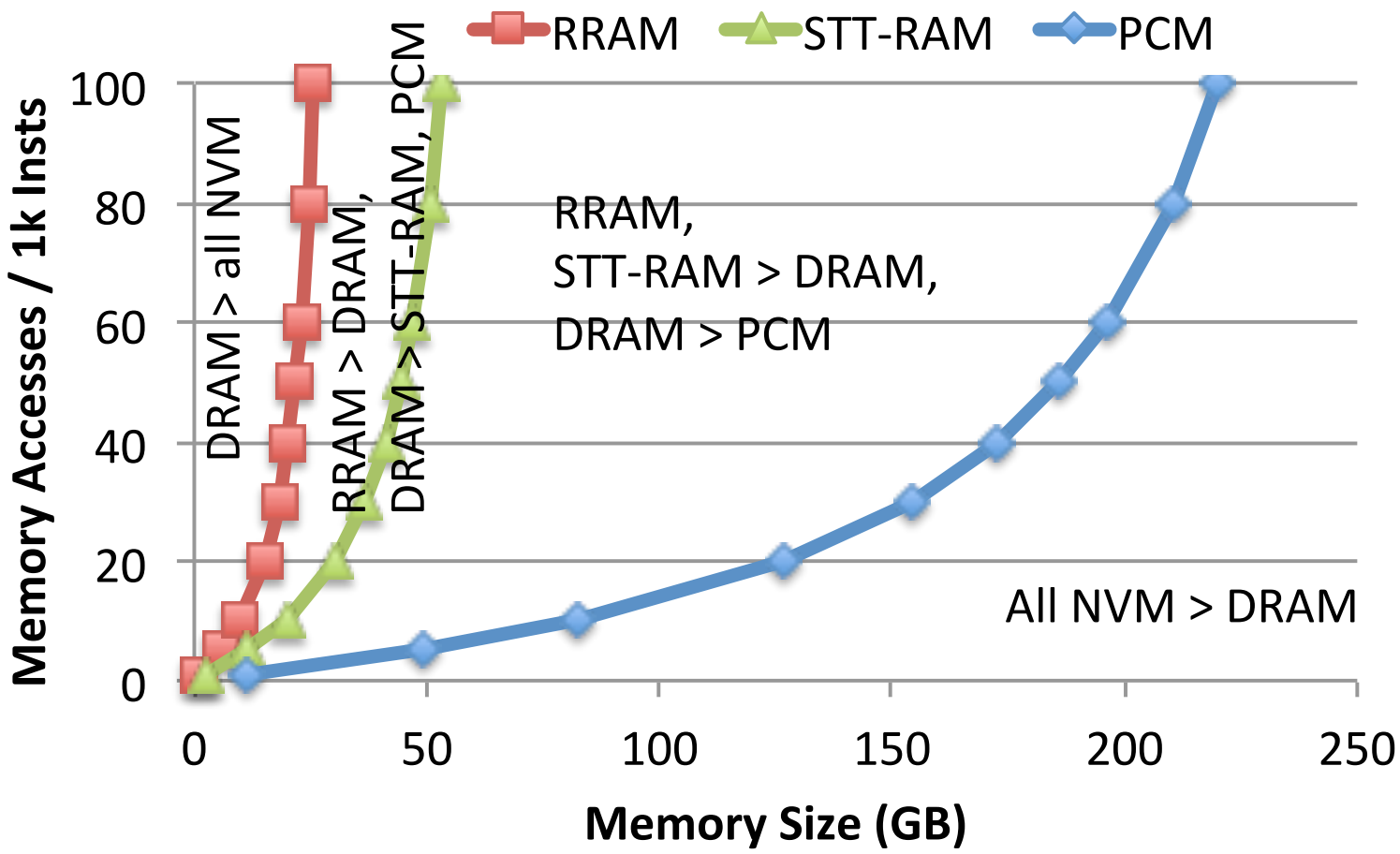
$$E_{mem} = E_{d,mem} NM + (P_{s,mem} S + P_{cpu}) T(L) \tag{5}$$

$$\Delta E = \frac{N}{\phi}(\phi\, \Delta E_d\, M + CPI_0\, \Delta P_s\, S + \Delta E_s\, M\, S + P_{cpu}\, M\, \Delta L) \tag{6}$$

[6] Hans Vandierendonck, Ahmad Hassan and Dimitrios S. Nikolopoulos: On the Energy-Efficiency of Byte-Addressable Non-Volatile Memory, *IEEE Computer Architecture Letters*, 2014.

NVRAM versus DRAM

Iso-energy-efficiency chart



Workload characterization for column stores

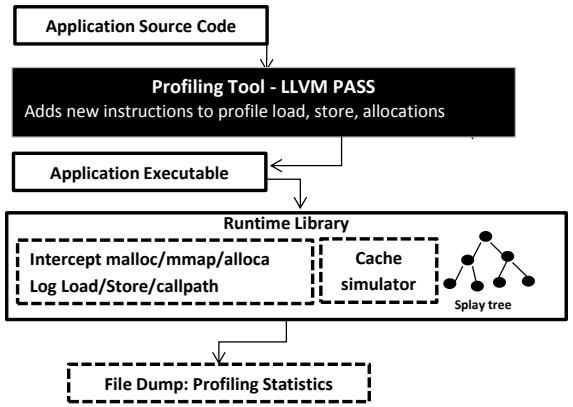


Figure : Object analysis tool

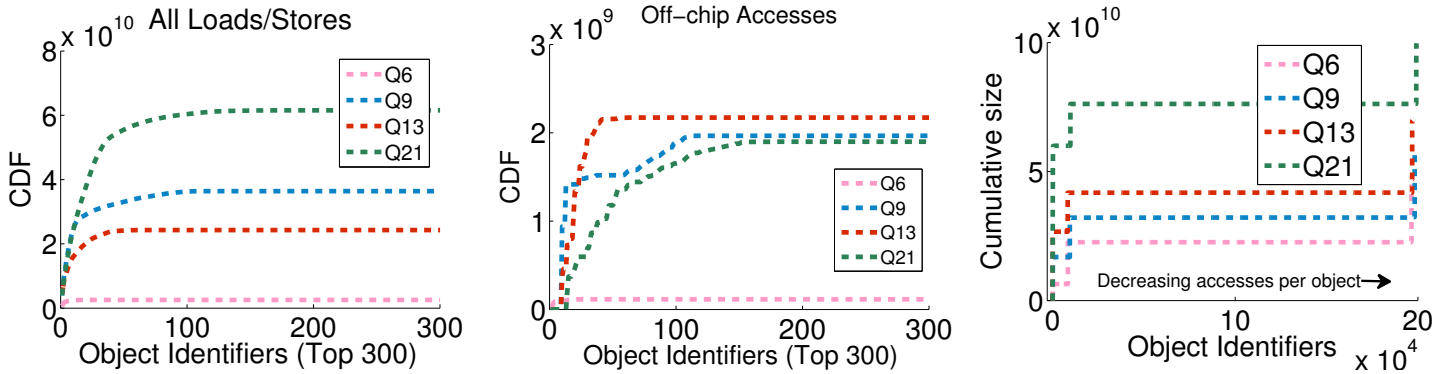


Figure : Workload Characterization of MonetDB.

Object placement in hybrid memories

$<$ 20% of objects needed in DRAM

Table : Device parameters

Hardware	**Specification**		
Server	Supermicro Intel(R) Xeon(R) CPU E5-4650, 2.70GHz, 32 cores, 20 MB LLC		
	Latency (cycles)	**Dynamic Energy (64 bytes)**	**Leakage Power**
DRAM	61 (R), 61 (W)	11.76 nJ(R), 25.35 nJ(W)	451 mW/GB
PCM	268 (R), 732 (W)	24 nJ(R), 1092 nJ(W)	4.23 mW/GB

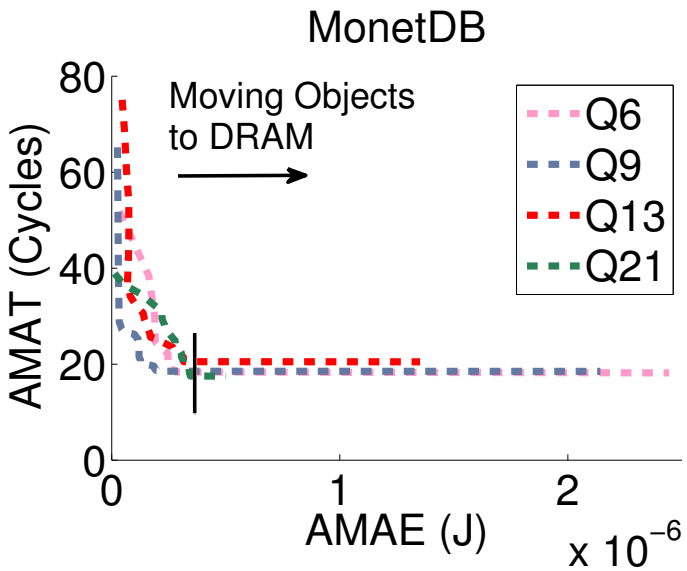


Figure : AMAT versus AMAE

Queen's University Belfast

Outline

1. HPC and the low-power processor ecosystem
2. The NanoStreams proposition
3. Financial real-time analytics
4. In-memory column stores
5. Conclusions

Queen's University Belfast

Where do we go from here

- Micro-server concept is not a stranger to HPC
  - BG/P and BG/Q would be good examples of state-of-the-art micro-servers for datacenters
- What could make it a value proposition
  - Improved energy-efficiency in applications where performance requirements are easily met
  - Improved energy-efficiency in data-intensive applications
  - Scale-out and tight-sizing machine for workload, rather than over-provision
- What may not be a value proposition
  - HPC applications that do require absolute peak performance
- What is needed
  - Holistic approaches: whole system design for energy-efficiency (memories, interconnect), co-designed software stack

Credits

EPSRC
Pioneering research and skills

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

- EU FP7 Grant 610509, EPSRC Grants L000055/1, L004232/1

- Charles Gillan, Giorgis Georgakoudis, George Tzenakis, Ahmad Hassan, Hans Vandierendonck, Bronis de Supinski