Server Resource Provisioning for Real-Time Analytics using Iso-Metrics

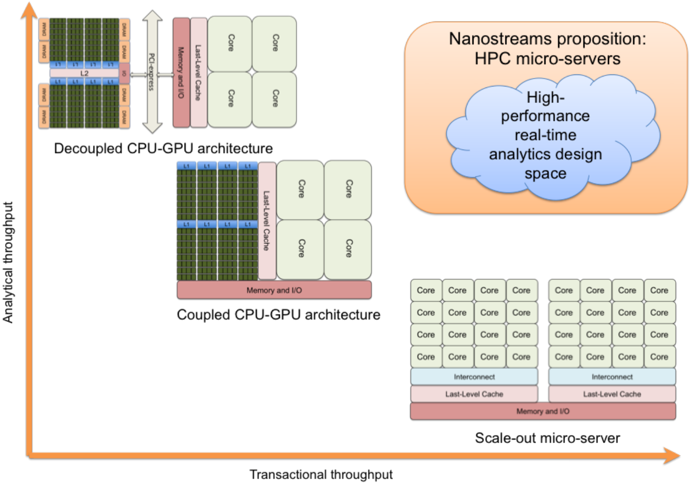
Dimitrios S. Nikolopoulos

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

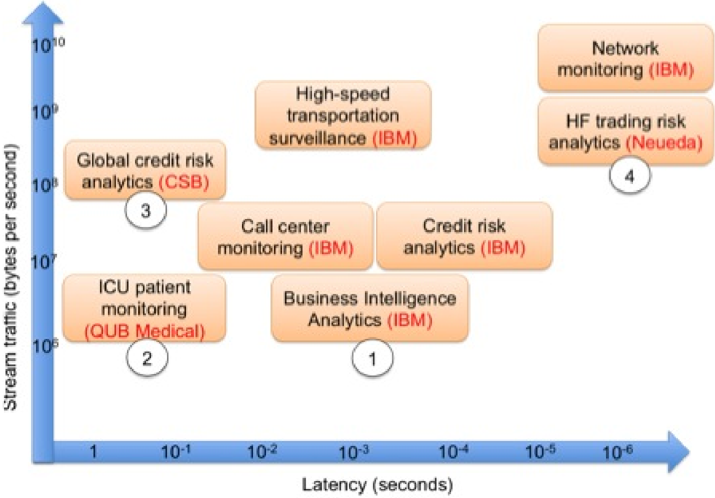
July 15, 2015

Outline

Diversity in the Server Landscape



Diversity in Workloads



How do we choose the right server?

QoS aware resource provisioning

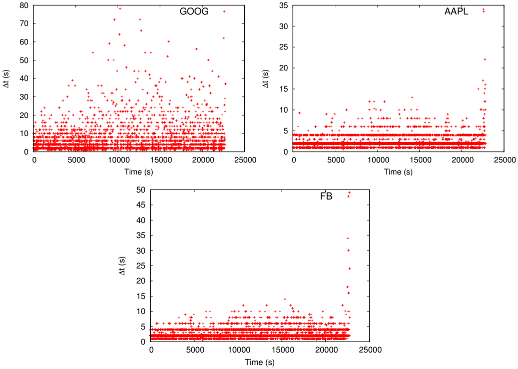
- Quality of Service (QoS)
  - End-user metric of performance
- Real-time analytics define QoS requirements
  - Online requests trigger analytic computations
  - User-driven, market-driven deadlines
- QoS aware resource provisioning
  - Energy savings from energy proportional execution
  - Better utilisation from multi-tenancy of services

Outline

Real-time option pricing

Georgakoudis et al. WHPCF'14, PPL

- Binomial Option Pricing: Black-Scholes generalisation in discrete time
- Symbol price change triggers options computation
- Deadline: compute *all* option contracts before next symbol change
- $QoS = \frac{Successful\ \ Pricings}{Total\ \ Pricings}$
- Inter-arrival time of price changes (Δt) varies per symbol



Option Pricing Kernels

Georgakoudis et al. WHPCF'14, PPL

- Not particularly compute- or data-intensive, low-latency and data-parallel workloads
  - Instance runs in ms or μs, must complete before next trade
  - Heavily traded symbols trigger Koptions/session

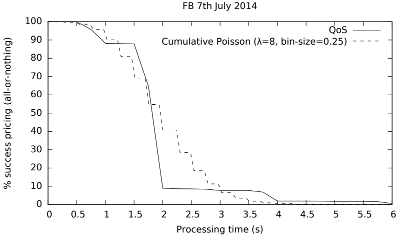
$$\mathrm{Price} = (-1)^p \left(SN((-1)^p d_1) - P\mathrm{e}^{-rT} N((-1)^p d_2) \right) \tag{1}$$

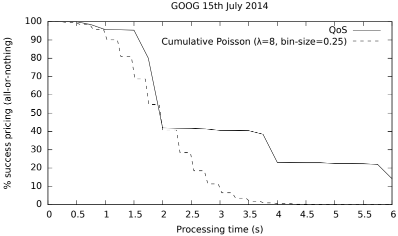
$$\mathrm{Price} = \frac{\mathrm{e}^{-rT}}{N} \sum_{i=1}^{N} \max \left(0, S - P\mathrm{e}^{(r - \frac{\sigma^2}{2})T + \sigma\sqrt{T} x_i} \right) \tag{2}$$

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

QoS in Detail

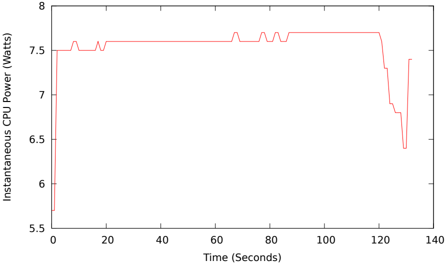
WHPCF'14, PPL





Cumulative frequency distribution of Facebook and Google stock price updates for full trading sessions on July 7th and 15th 2014

Iso-QoS and Energy



$$\text{QoS(t)} = 1 - e^{-\lambda} \sum_{i=0}^{t} \frac{\lambda^t}{\lfloor t! \rfloor} \quad (4)$$

$$G \geq N_{\text{opt}} \times S_{\text{opt}} \quad (5)$$

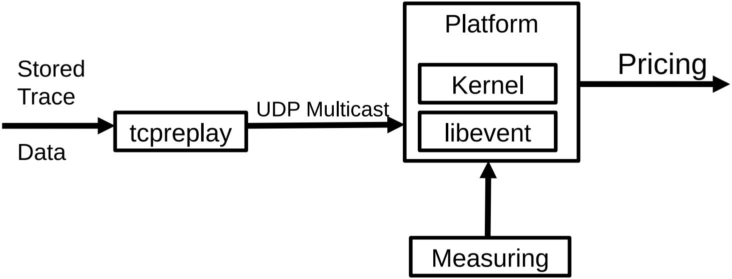
$$E_{\text{gap}} = N_{\text{opt}} \times J_{\text{opt}} \quad (6)$$

$$N_{\text{gaps}} = \lfloor Y \times \text{session\ \ updates)} \rfloor \quad (7)$$

$$E_{\text{QoS=Y}} = N_{\text{gaps}} \times E_{\text{gap}} \quad (8)$$

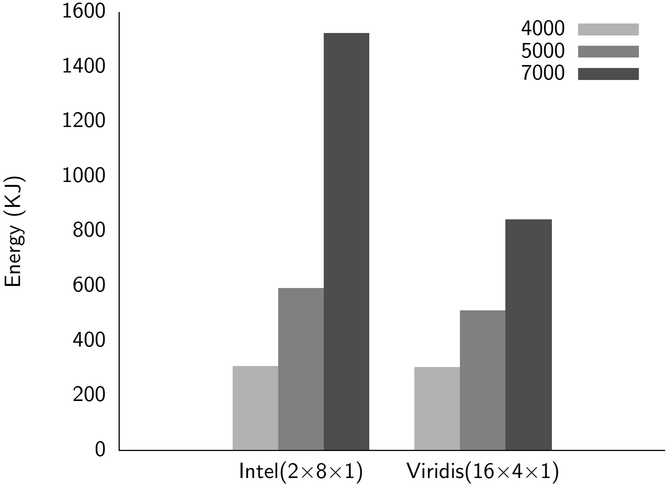
- **Joules/option**: Provider-side, sustained throughout trading day, reduction translates to less TCO
- **Time/option**: User-side, end-to-end latency.
- **QoS:** Calculating option before new price arrives; unknown deadline.

Feed Handling



Financial trace data measurement setup

Servers vs. Micro-Servers



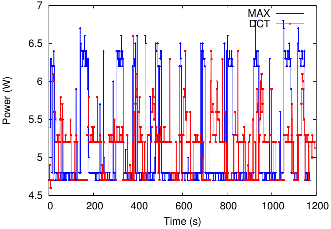
BT kernel energy consumption scaling (at QoS=80%) of Viridis(16×4×1) and Intel(2×8×1)

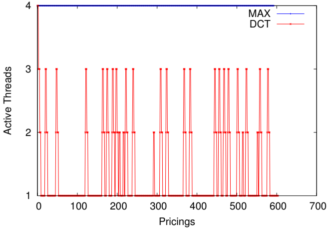
Qos aware dynamic concurrency throttling

- Supervisor with a QoS target manages execution
  - Receive QoS feedback from pricers
  - Send number of computation threads to pricers
- Simple Feedback-driven DCT policy
  - $+1$ thread (up to maximum), if $QoS < QoS_{target}$
  - -1 thread, if $QoS > (QoS_{target} + Threshold)$

Results

- Experiment on a micro-server node, 4 $\times$ Cortex A9, 4GB RAM
- Compare MAX policy (maximum number of threads) vs. QoS-aware DCT policy
- DCT $QoS = 92\%$, 15% less peak power consumption, fewer active threads





Outline

New low-power computing paradigms

Nikolopulos et al. IEEE Computer

- Approximate computation:
  - Deliberately drop computation
  - Effective in signal processing algorithms
  - Active research in language, compiler, runtime, architecture support
- Significance-driven computation:
  - Generalisation of approximate computing
  - Invests more resources in the efficient execution of most significant instructions

Case study: self-stabilising CG

EEHCO'15, IET CDT

Algorithmic based fault correction

- Periodic step that corrects the state of the algorithm
- Convergence if healing step accurate
- No assumption about convergence rate

Hybrid architecture - $A15 + NA7$

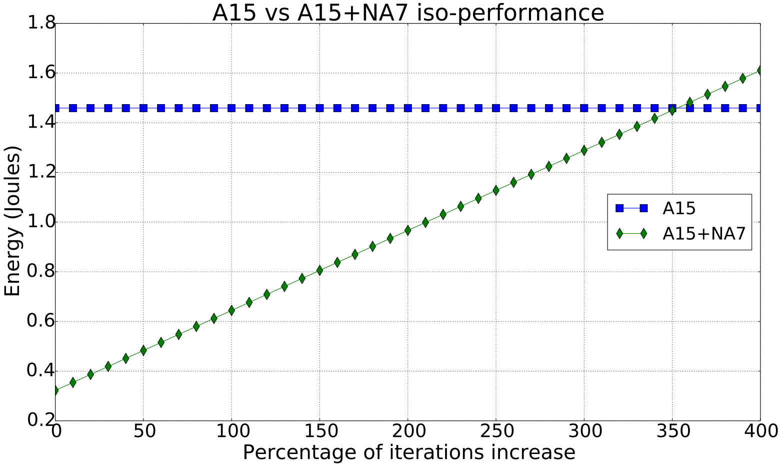
- One reliable $A15$, N unreliable $A7$
- Healing step on $A15$ core, normal steps on $A7$ cores

case study	#A7 clusters	GFLOPS	Power
iso-performance	5.51	2.09	1.24
iso-power	38.85	13.49	5.44
iso-capacity	4	1.57	1.05

Evaluation of performance, power and energy on the target architectures

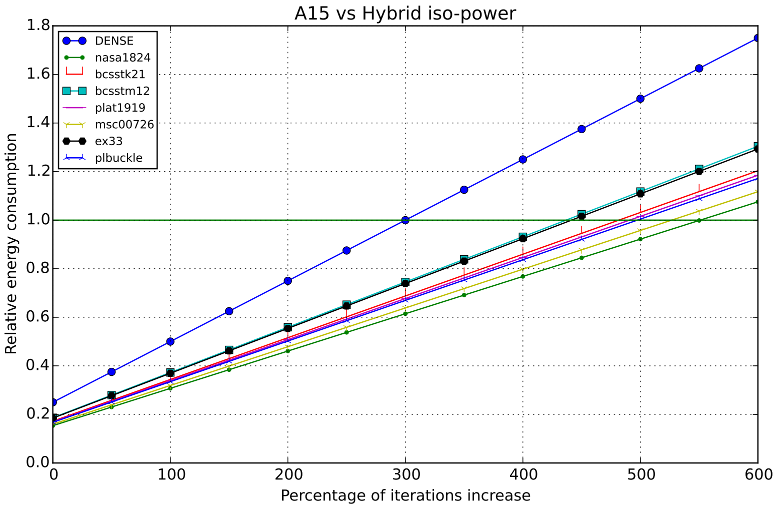
Case study: self-stabilising CG

EEHCO'15, IET CDT



Case study: fault-tolerant GMRES

EEHCO'15, IET CDT



Outline

Conclusion

- Efficiency best understood through platform-agnostic metrics and methodologies
- Iso-metrics provide fair ranking of algorithms, architectures, systems
- Mathematical formulation of QoS metrics for real-time analytics
- Iso-QoS produces surprising server rankings for real-time applications
- New low power heterogeneous computing paradigms driven by iso-metrics

Queen's University Belfast

Credits

nanostreams

SCoRPiO
Significance-Based Computing for
Reliability and Power Optimization

SEVENTH FRAMEWORK PROGRAMME

EPSRC
Pioneering research and skills

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

SEVENTH FRAMEWORK PROGRAMME

- EU (Grants: 323872, 610509), EPSRC (Grants: L000055/1, L004232/1)
- Collaborators: Kirk Cameron, Sandra Catalán, Charalambos Chalios, Giorgis Georgakoudis, Charles Gillan, Ahmad Hassan, George Karakonstantis, Dong Li, Enrique Quintana Ortií, Martin Schulz, Ivor Spence, Bronis R. de Supinski, George Tzenakis, Hans Vandierendonck, Vassilis Papaefstathiou, and the NanoStreams, SCoRPiO, ALEA, ENPOWER project consortia.