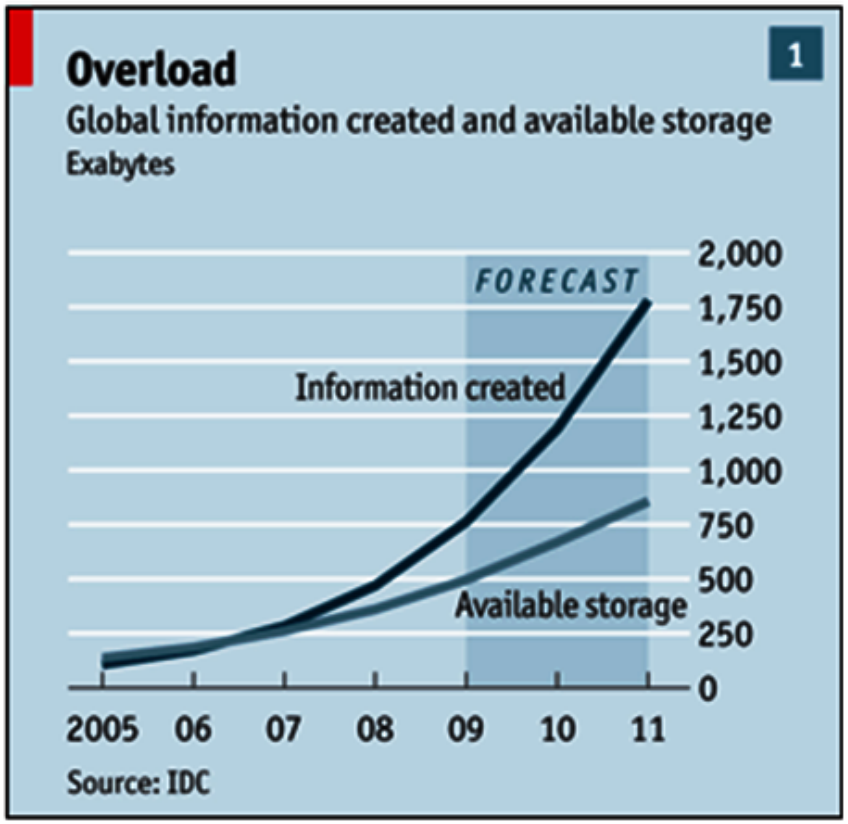
Mega-servers vs Micro-blades for Data Centers

Benjamin C. Lee
Stanford University
bcclee@stanford.edu

Architectural Concerns in Large Datacenters
International Symposium on Computer Architecture
19 June 2010

Data Deluge

- Data centers are not keeping up
- Data is not information!



"Data, data everywhere." *The Economist*, 25 Feb 2010.

Value from Data

- **Statistical Inference**
  - Draw useful information from free data
  - Useful information applies statistical inference
  - Free data drawn from Internet's webpages

- **Ex: Langugage Translation**
  - Statistical inference trumps linguistic structure
  - Early 1990's, IBM French/English with O(1E+6) documents
  - Presently, Google Translate with O(1E+9) documents

"Clicking for gold." *The Economist*, 25 Feb 2010.

Value from Data

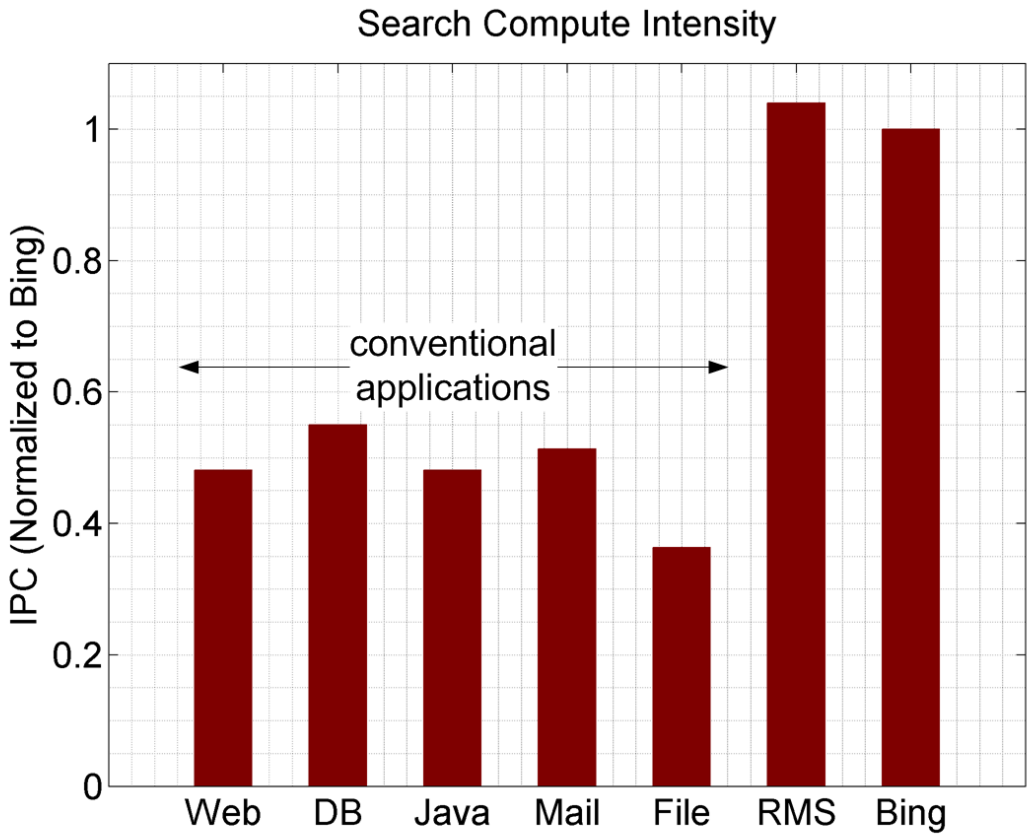
- **Statistical Inference**
  - Draw useful information from free data
  - Useful information applies statistical inference
  - Free data drawn from Internet's webpages

- **Ex: Langugage Translation**
  - Statistical inference trumps linguistic structure
  - Early 1990's, IBM French/English with O(1E+6) documents
  - Presently, Google Translate with O(1E+9) documents

"Clicking for gold." *The Economist*, 25 Feb 2010.

Computational Intensity

- Bing web search ranks pages with neural network
- RMS foreshadows future analytic workloads



V.J. Reddi et al. "Web search using mobile cores" *ISCA*, 2010.

SW: Latency is Paramount

- **Applications are Evolving**
  - Apps not memory, I/O bound anymore!
  - Past: Response time insensitive to compute latency
  - Present: Response quality sensitive to compute latency
  - Future: Algorithmic gains require expendable latency

- **Architectures are Conservative**
  - Separation of SW/HW interests
  - Deploy homogeneous, over-provisioned hardware

SW: Latency is Paramount

- **Applications are Evolving**
  - Apps not memory, I/O bound anymore!
  - Past: Response time insensitive to compute latency
  - Present: Response quality sensitive to compute latency
  - Future: Algorithmic gains require expendable latency
- **Architectures are Conservative**
  - Separation of SW/HW interests
  - Deploy homogeneous, over-provisioned hardware

HW: Efficiency is Paramount

- **Multicore is Unsustainable**
  - Dennard scaling is dead!
  - Past: Dennard provides constant power density
  - Present: Scaling increases power density
  - Future: Many-core unrealizable

- **Heterogeneity is Solution**
  - Go big or go home!
  - Past: CMP of homogeneous cores
  - Present: CMP of big/small & graphics cores
  - Future: SoC of general-purpose & accelerators

HW: Efficiency is Paramount

- **Multicore is Unsustainable**
  - Dennard scaling is dead!
  - Past: Dennard provides constant power density
  - Present: Scaling increases power density
  - Future: Many-core unrealizable

- **Heterogeneity is Solution**
  - Go big or go home!
  - Past: CMP of homogeneous cores
  - Present: CMP of big/small & graphics cores
  - Future: SoC of general-purpose & accelerators

Economics is Paramount

- **Specialization is Expensive**
  - O(10M)$ for custom HW/SW!
  - Past: Customization for standard computation
  - Present: ASIC production falling
  - Future: Generalizable specialization

- **Economics can Improve**
  - Increase volume via generality, more app targets
  - Reduce cost via methodology
  - Technology challenges even more costly!

Economics is Paramount

- **Specialization is Expensive**
  - O(10M)$ for custom HW/SW!
  - Past: Customization for standard computation
  - Present: ASIC production falling
  - Future: Generalizable specialization

- **Economics can Improve**
  - Increase volume via generality, more app targets
  - Reduce cost via methodology
  - Technology challenges even more costly!

Mega-servers vs Micro-blades for Data Centers

Benjamin C. Lee
Stanford University
bcclee@stanford.edu

Architectural Concerns in Large Datacenters
International Symposium on Computer Architecture
19 June 2010