Scaling Challenges for Warehouse Scale Computers

Bikash Koley
Lead, Network Architecture and Capacity Planning

April 26, 2011
Santa Barbara Summit on Energy Efficiency
Google’s Mission

To organize the world's information and make it **UNIVERSALLY** accessible and useful.
• **User base**
  - World population: 6.676 billion people (June'08, US Census est.)
  - Internet Users: 1.463 billion (>20%) (June'08, Nielsen/ITU)
  - Google Search: More than a billion searches daily

• **Geographical Distribution**
  - Google services are worldwide: over 55 countries and 112 languages
  - More than half our searches come from outside the U.S.

• **Data Growth**
  - Web expands/changes: billions of new/modified pages every month
  - Every few hours we crawl/refresh more than whole Library of Congress
  - YouTube gains over 13-15 48 24 hours of video every minute, 1+ billion views a day

• **Latency Challenge**
  - Speed of Light in glass: $2 \times 10^8 \text{ m/s} = 2,000 \text{ km} / 10 \text{ ms}$
  - “Blink of an eye response” = 100 ms
Warehouse Scale Computers

Consolidated Computing, Many UIs, Many Apps, Many Locations

Luiz André Barroso, Urs Hölzle, “The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines”,
http://www.morganclaypool.com/doi/abs/10.2200/S00193ED1V01Y200905CAC006?prevSearch=allfield%253A%2528Urs%2529&searchHistoryKey=
Warehouse-size Computer

Warehouse-scale Computer
Why not just call it a data center?

Data center

- Co-located machines that share security, environmental requirements
- Applications == a few binaries, running on a small number of machines
- Heterogeneous hardware and system software
- Partitioned resources, managed and scheduled separately
- Facility and computing equipment designed separately

Warehouse-scale computer

- Computing system designed to run massive Internet services
- Applications == tens of binaries running on hundreds-thousands of machines
- Homogeneous hardware and system software
- Common pool of resources managed centrally
- Integrated design of facility and computing machinery
Energy Efficient Data Centers

\[ PUE = \frac{Total\_Facility\_Power}{IT\_Equipment\_Power} \]

Industry Avg PUE: \(~ 1.9^*\)

Google Average PUE for the last 12 months: < 1.16

Environmental Protection Agency, ENERGY STAR Program, 2007, “Report to Congress on Server and Data Energy Efficiency,”
Efficient Data Centres

• Cooling is 30-70% of overhead
• Evaporative cooling enables “free cooling” (no chiller usage)
• But fresh water is valuable
• Use recycled water instead...
Efficient Servers

- High-quality Power Supplies
- Very efficient voltage regulator modules
- Omit irrelevant components (e.g. graphics chips)
- Minimize fan power
- Goal: “energy proportionality”
Hardware & Energy Efficiency

• Moore’s law is now scaling # cores instead of MHz
  – Fine with us: we love multicore machines for our problems

• Still want more computing capabilities, though...
  – Easy to get more computation by using more energy
  – Proportion of costs for energy will continue to grow, since Moore’s law keeps computing cost roughly fixed

• Challenge: for every increase in HW performance, we need a corresponding increase in energy efficiency
Networking and Energy Efficiency

– Large number of identical compute systems
– Interconnected by a large number of identical switching gears
– Can be within single physical boundary or can span several physical boundaries
– Interconnect length varies between few meters to hundreds of kms

What we would prefer:

Today

Fiber-rich

Fiber-scarce

Available BW

Distance Between Compute Elements
• High performance computing/super-computing architectures have often used various complex multi-stage fabric architectures such as Clos, Fat Tree or Torus
• For this theoretical analysis, we picked the Fat Tree architecture described in [2, 3]

For a $k$ stage fabric with $N$-port nodes and $N$ leaf nodes

$$System\_Port\_Cnt = \frac{N^k}{2^{k-1}}$$

$$System\_Node\_Count = N^{k-1}\left(1 + \frac{1}{2^{k-1}}\right)$$
Interconnect at What Port Speed?

- A switching node has a fixed switching capacity (i.e. CMOS gate-count) within the same space and power envelope
  - Per node switching capacity can be presented at different port-speed:
    - i.e. a 400Gbps node can be 40X10Gbps or 10X40Gbps or 4X100Gbps
    - Lower per-port speed allows building a much larger size maximal constant bisectional bandwidth fabric
    - There are of course trade-offs with the number of fiber-connections needed to build the interconnect
    - Higher port-speed may allow better utilization of the fabric capacity
Power vs Port Speed

- Three power consumption curves for interface optical modules:
  - Bleeding Edge: 20x power for 10x speed; e.g. if 10G is 1W/port, 100G is 20W/port
  - Power Parity: Power parity on per Gbps basis; e.g. if 10G is 1W/port, 100G is 10W/port
  - Mature: 4x power for 10x speed; e.g. if 10G is 1W/port, 100G is 4W/port
  - Lower port speed provides lower power consumption
  - For power consumption parity, power per optical module needs to follow the “mature” curve
Thank You!