Energy-Efficiency & Large-Scale Computing: Networking's Role

Prof. Margaret Martonosi
Princeton University
mrm@princeton.edu

PRINCETON

Aggregate IT Energy Consumption

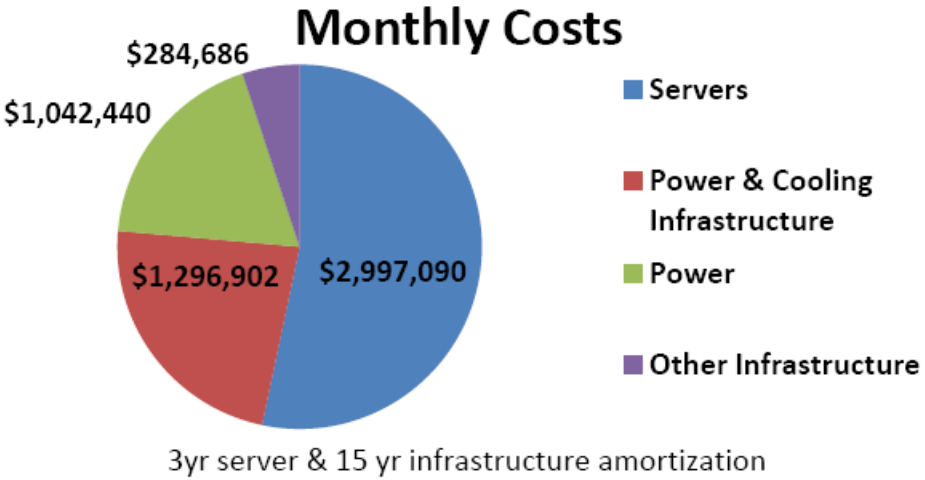
- Information (and communications) Technology (IT) consumes 2.5% of the world's electricity
  = 1B tons of CO2 annually.
- In the US, data centers alone consume more than 60B KWH per year
  = energy consumed by entire transportation manufacturing sector.
- Current trends: energy usage will nearly double by 2011 for overall electricity cost of $7.4 B per year.

Kyoto Protocol

- Sets binding targets for 37 industrialized countries and the EC for reducing greenhouse gas (GHG) emissions
  - Avg 5% against 1990 levels over the five-year period 2008-2012
- Adopted in late 1997 & entered into force on 2/16/2005. 182 parties/countries ratified so far…
- Three market-based mechanisms
  - Emissions trading – known as “the carbon market”
  - the clean development mechanism (CDM)
  - joint implementation (JI).

Data Center Costs

- **Assumptions:**
  - Facility: ~$200M for 15MW facility (15-year amort.)
  - Servers: ~$2k/each, roughly 50,000 (3-year amort.)
  - Average server power draw at 30% utilization: 80%
  - Commercial Power: ~$0.07/kWhr



3yr server & 15 yr infrastructure amortization

- **Observations:**
  - $2.3M/month from charges functionally related to power
  - Power related costs trending flat or up while server costs trending down

Details at: http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx

(James Hamilton, ISCA '09 keynote)

PRINCETON

Data Center Energy Optimizations

- Optimize data center operation with respect to:
  - Power/performance tradeoffs
  - Electricity costs
  - Electricity type (“green” production or carbon-based)
- All subject to performance goals in SLA
- Main Control Mechanisms:
  - Request Distribution across geographically-distributed data centers
  - Dynamic power management and server control within data centers
    - Typically not turn on/off, but rather migration and local request distribution

Network opportunities

- If turning servers on/off isn't a good option, then best control approaches are through routing, migration, request distribution
- Network Latency:
  - Predictability
  - Observability
  - Low-latency for jobs that require it (not all)
- Network Observability:
  - Measure latency
  - Count event types
  - Collect and analyze information for informed adaptive control