Internet-scale Computing: The Berkeley RADLab Perspective

IWQoS 2007
Evanston, IL

Randy H. Katz

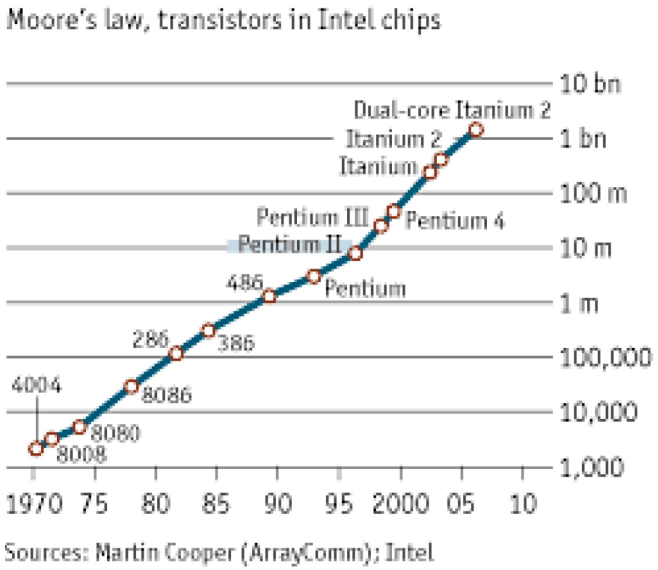
randy@cs.berkeley.edu

21 June 2007

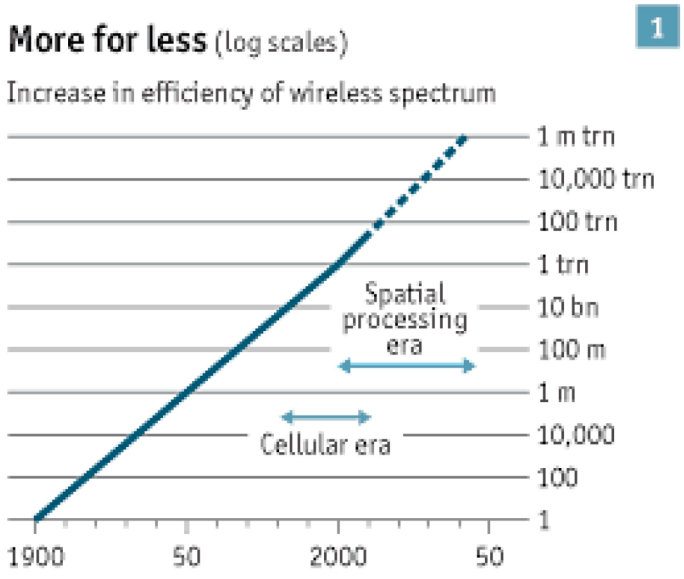
The Challenge for IWQoS

- QoS in a world of cheap processing and storage, and plentiful network b/w?
- Oh, but network b/w is limited … Or is it?

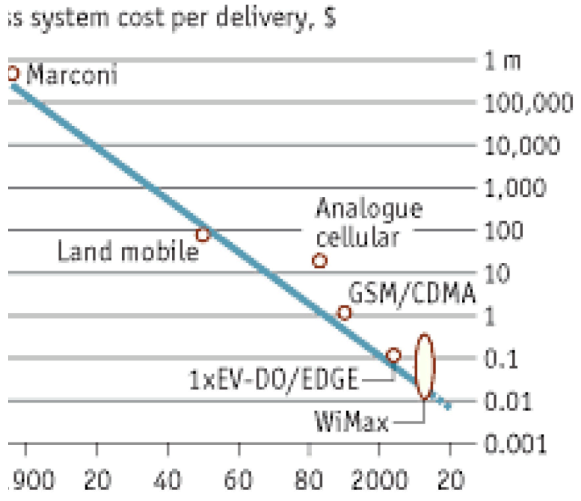
Rapidly increasing transistor density



"Spectral Efficiency": More bits/m^3



Rapidly declining system cost



Not So Simple …

Speaking in tongues 2

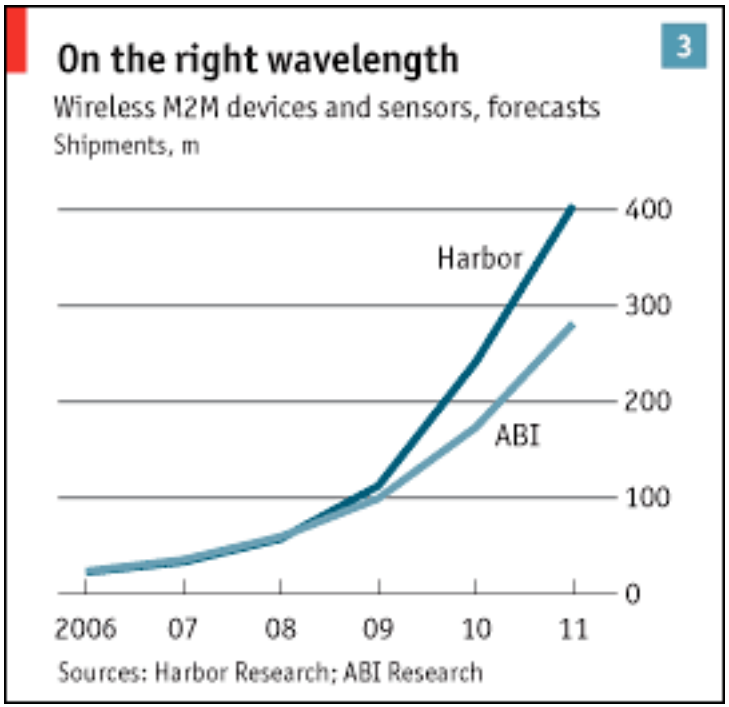
Main two-way wireless technologies*

	Data rate per second	Range	Cost†
Mobile WiMax	15Mb	5km	$8 in 2008
3G cellular (HSDPA/LTE)	14Mb	10km	$6
2G cellular (GSM/CDMA)	400k	35km	$5
Wi-Fi	54Mb	50-100m	$4
Bluetooth	700k	10m	$1
Zigbee	250k	30m	$4
UWB	~400Mb	5-10m	$5
RFID	1-200k	0.01-10m	4 cents

*Typical performance; actual figures vary
†Approx. device-chip cost at high volume

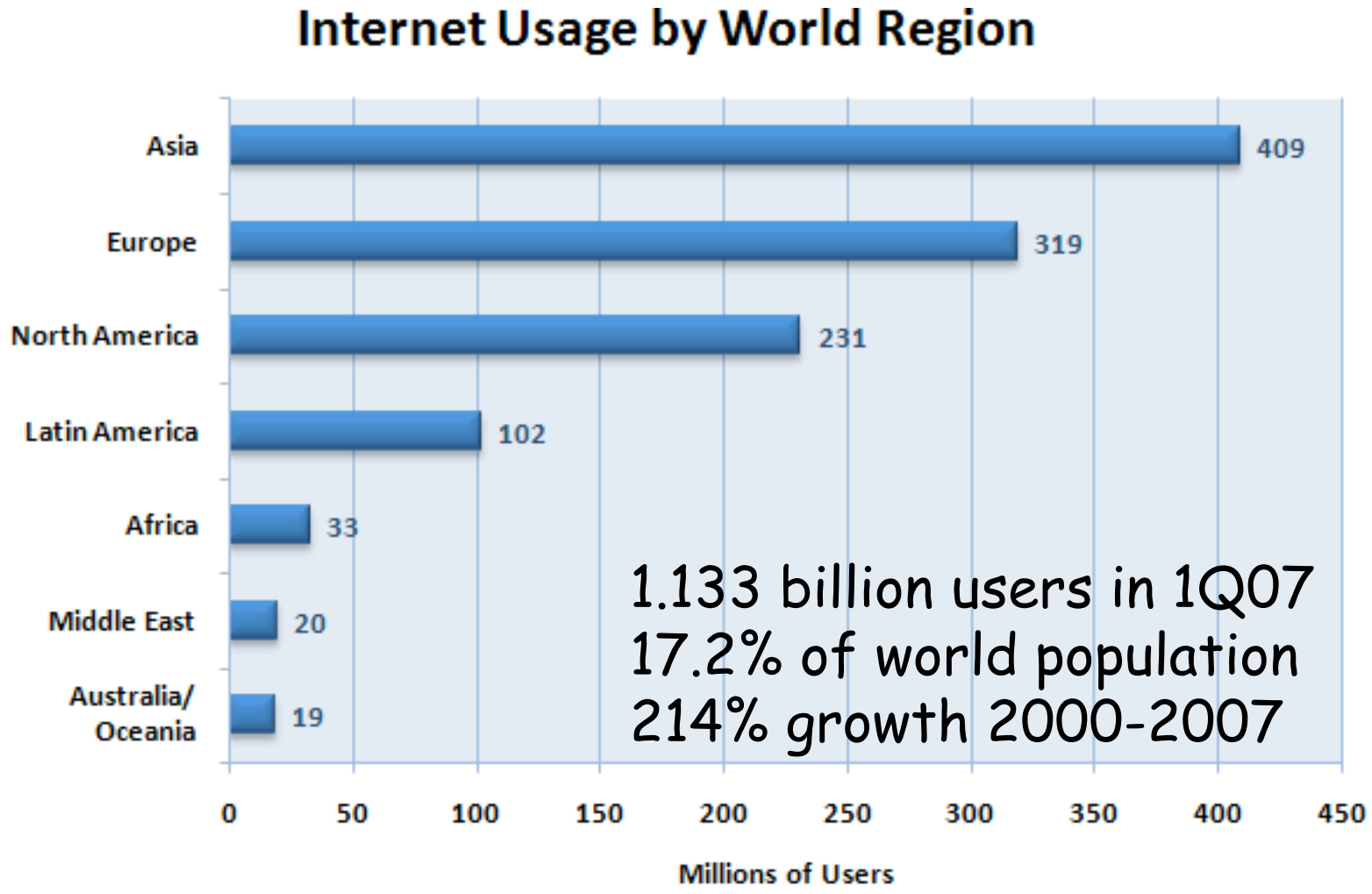
Sources: William Webb; Cambridge Consultants; OECD; Pyramid Research; Nokia; TI; CSR; Ember; Hitachi

Speed-Distance-Cost Tradeoffs



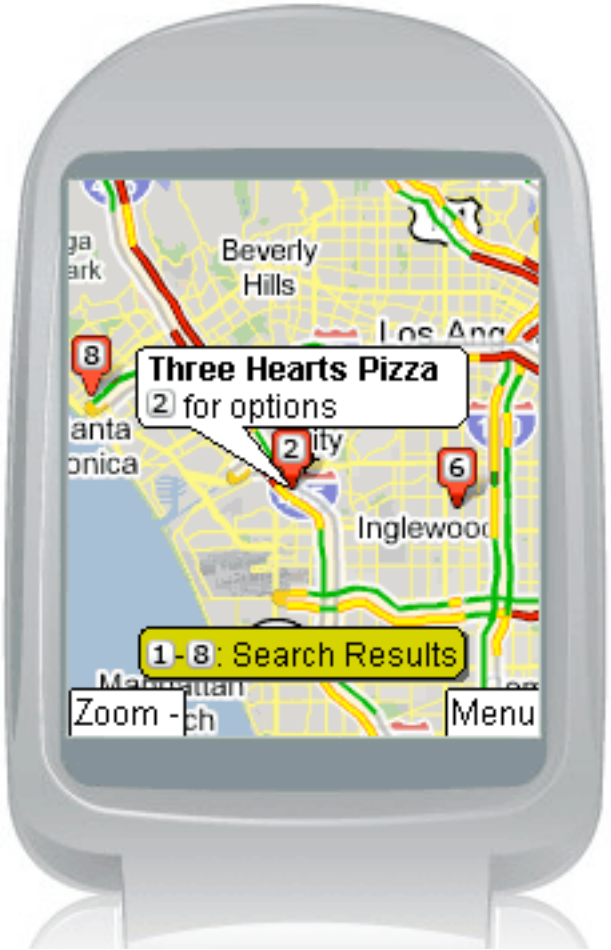
Rapid Growth: Machine-to-Machine Devices (mostly sensors)

Sensors Interesting, But …



But What about Mobile Subscribers?

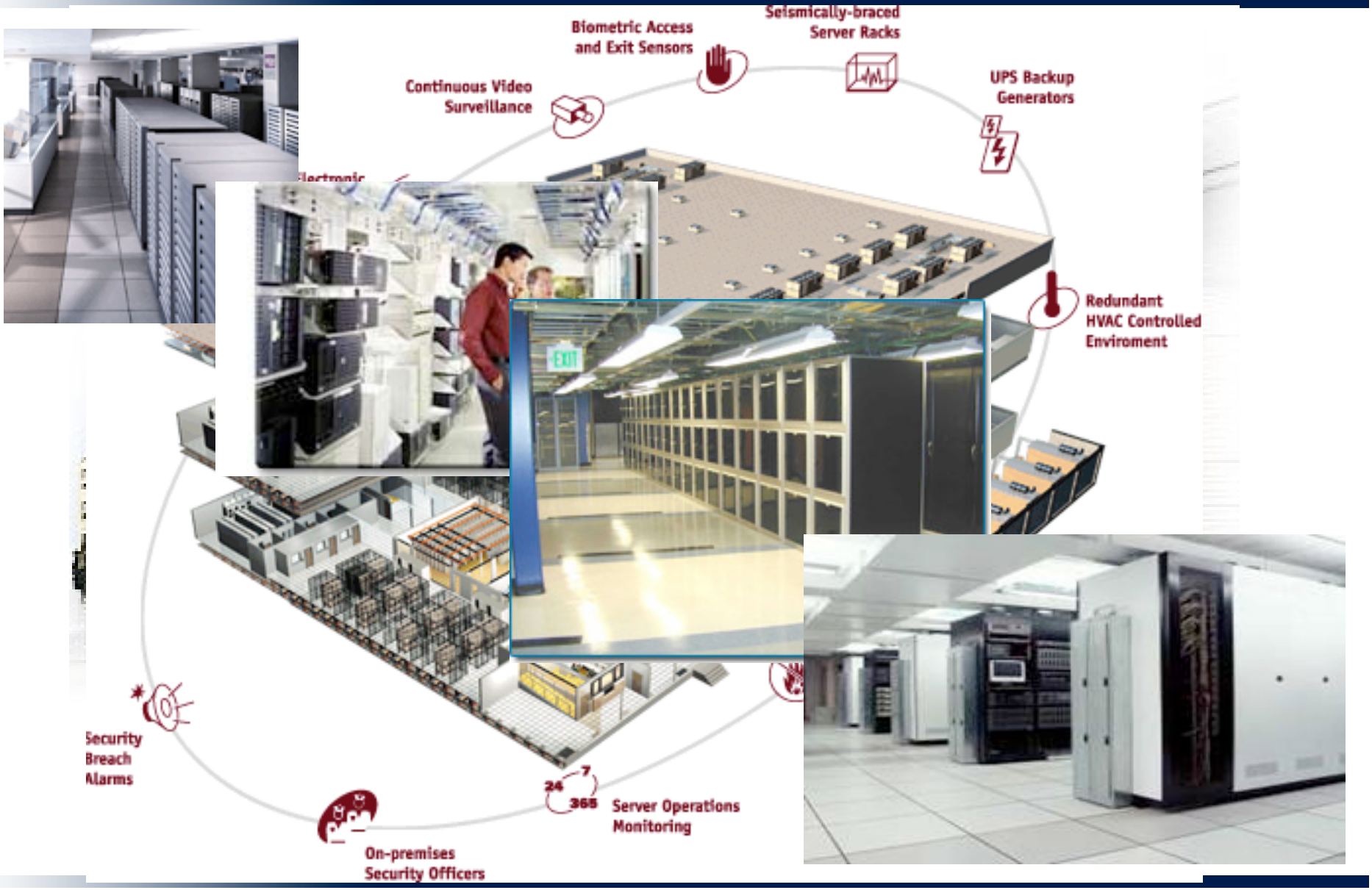
These are Actually Network-Connected Computers!



2007 Announcements by Microsoft and Google

- Microsoft and Google race to build next-gen DCs
  - Microsoft announces a $550 million DC in TX
  - Google confirm plans for a $600 million site in NC
  - Google two more DCs in SC; may cost another $950 million -- about 150,000 computers each
- Internet DCs are the next computing platform
- Power availability drives deployment decisions

Internet Datacenters as Essential Net Infrastructure



Keep on trucking

AMERICAN POWER CONVERSION CORP.'S InfraStruxure Express mobile data center can deliver power and Internet connectivity when there are no other options.

InfraStruxure Express is a fully operational data center on wheels (see photo)

officials said that the cost of a lease depends on financing options but that companies could expect to pay about $20,000 per month. They added that InfraStruxure Express can be delivered anywhere in the continental United States within three and a half days

provide as much as 400 kilowatts of power, and it has external feeds that can be used to deliver temporary power to buildings.

The on-board cooling is adequate for data center environments, and the trailer is configured with hot and cold aisles for

Datacenter is the Computer

- Google *program* == Web search, Gmail,…
- Google *computer* ==

  Warehouse-sized facilities and workloads likely more common

  Luiz Barroso's talk at RAD Lab 12/11/06

Sun Project Blackbox 10/17/06

Compose datacenter from 20 ft. containers!

- Power/cooling for 200 KW
- External taps for electricity, network, cold water
- 250 Servers, 7 TB DRAM, or 1.5 PB disk in 2006
- 20% energy savings
- 1/10th? cost of a building

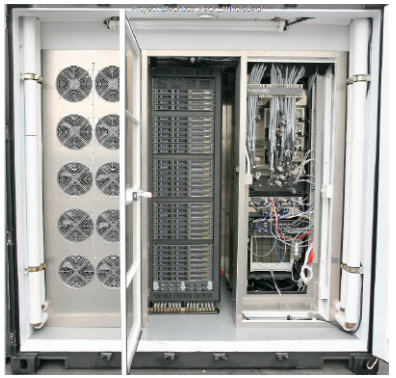
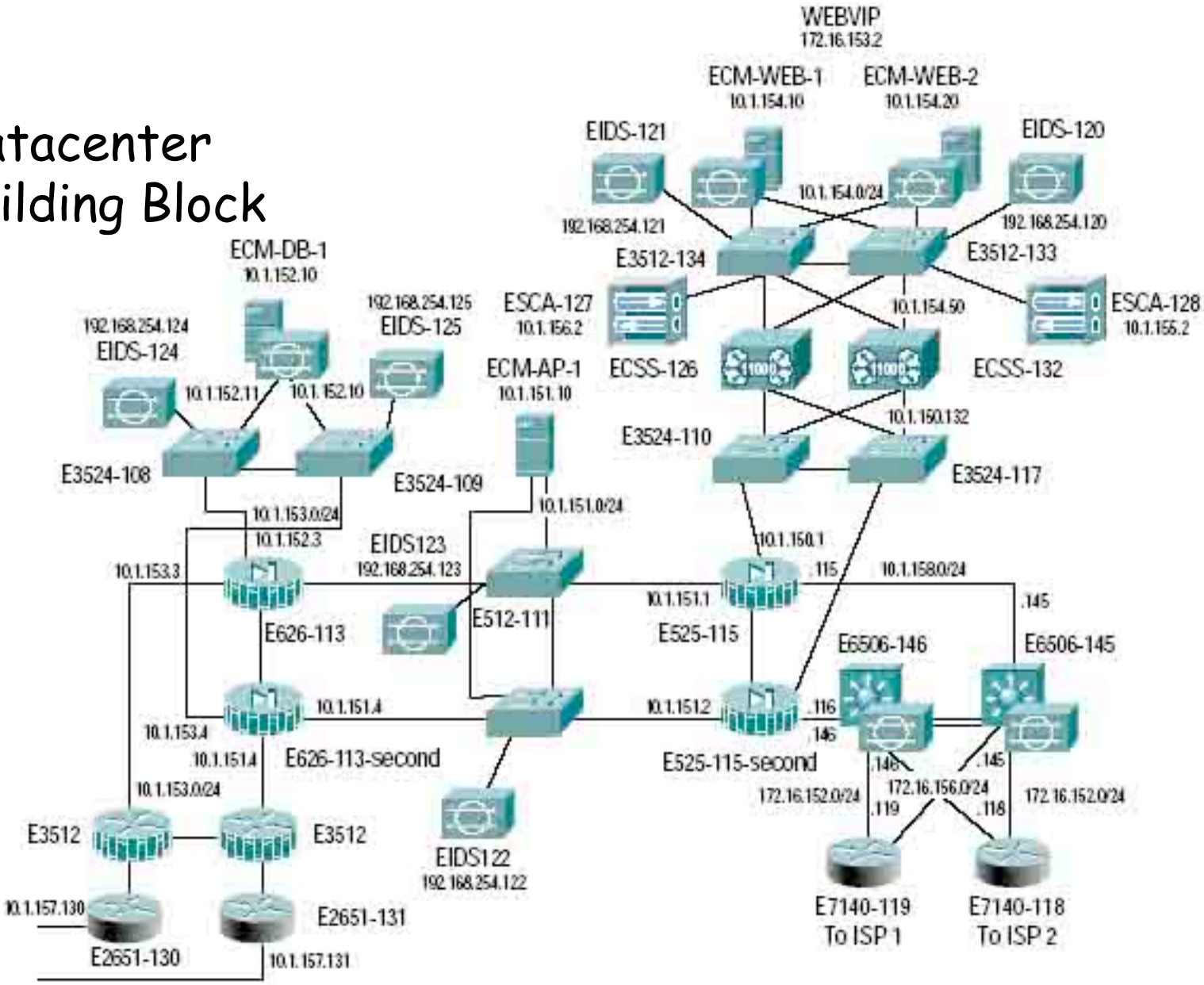
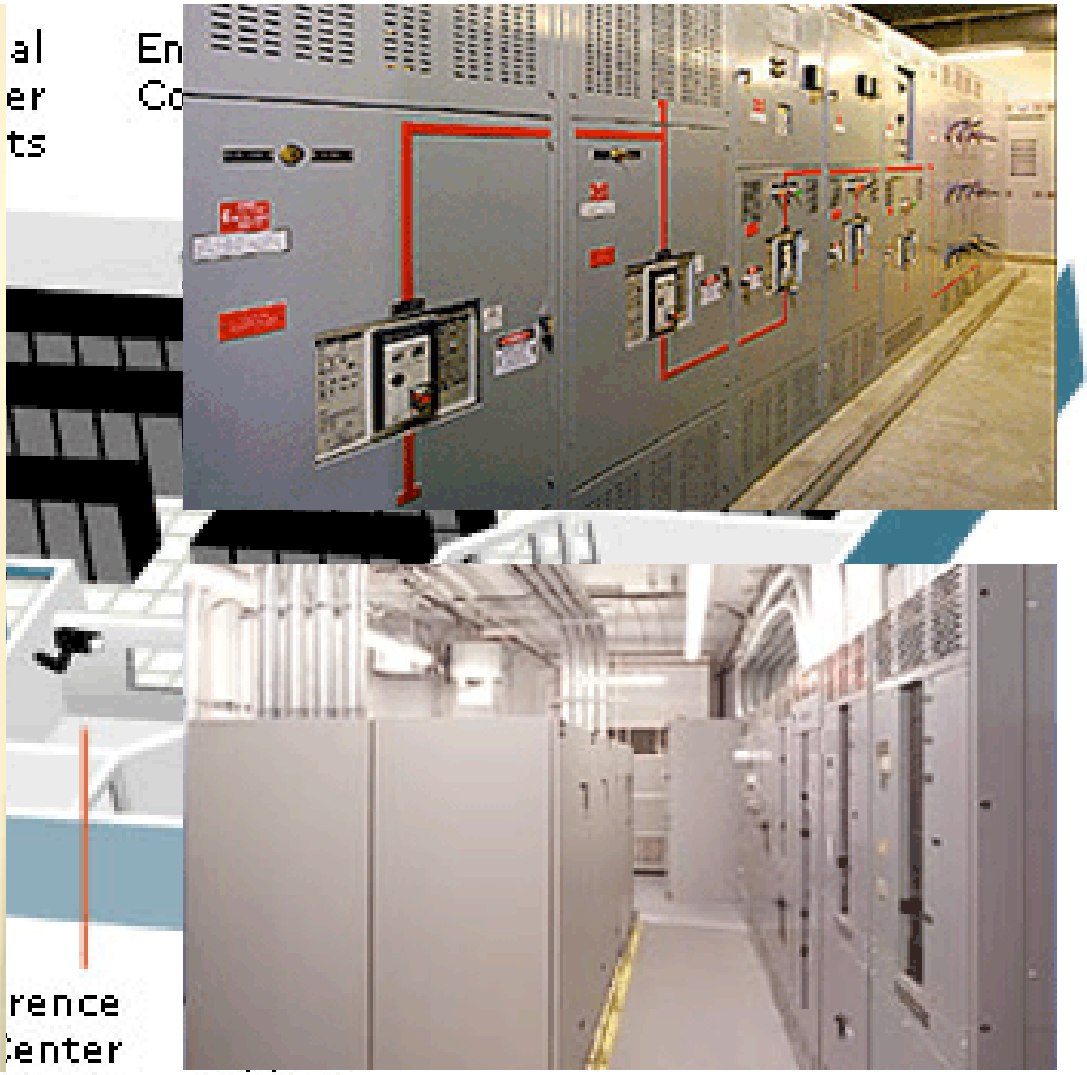


Figure 40
E-Commerce/Data Center Module: Detail

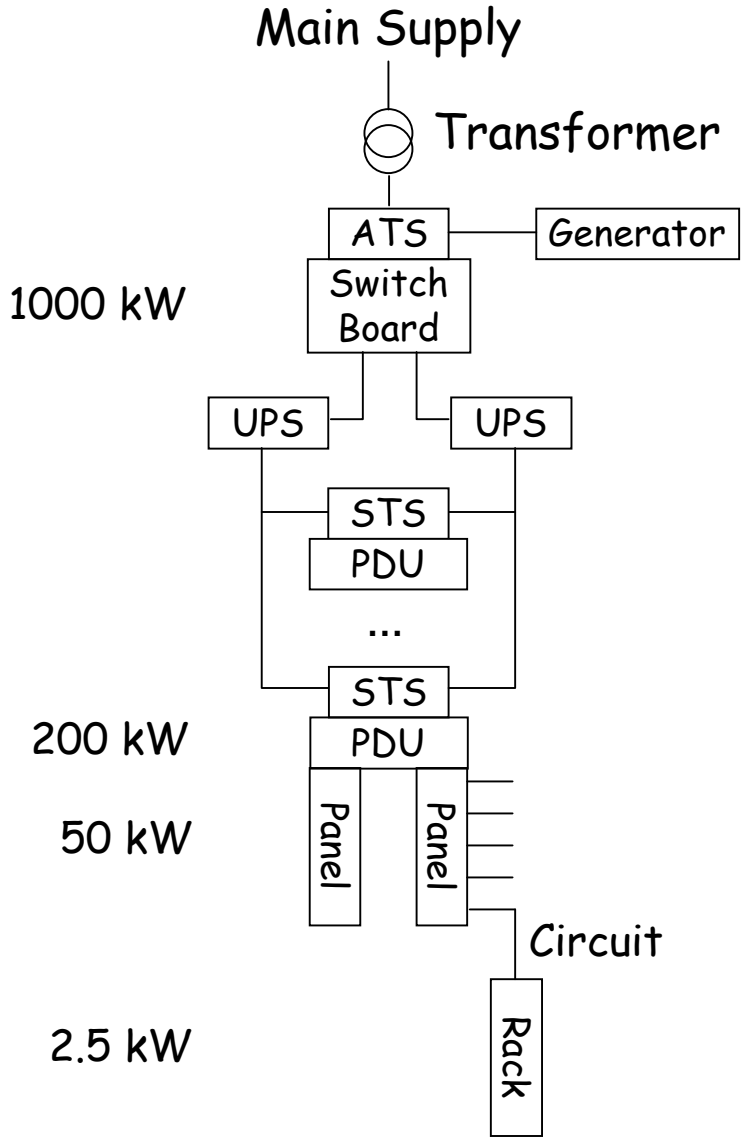
"Typical" Datacenter Network Building Block



Computers + Net + Storage + *Power* + *Cooling*



Datacenter Power Issues



- Typical structure 1MW Tier-2 datacenter
- Reliable Power
  - Mains + Generator
  - Dual UPS
- Units of Aggregation
  - Rack (10-80 nodes)
  - PDU (20-60 racks)
  - Facility/Datacenter

X. Fan, W-D Weber, L. Barroso, "Power Provisioning for a Warehouse-sized Computer," ISCA'07, San Diego, (June 2007).

Nameplate vs. Actual Peak

Component	Peak Power	Count	Total
CPU	40 W	2	80 W
Memory	9 W	4	36 W
Disk	12 W	1	12 W
PCI Slots	25 W	2	50 W
Mother Board	25 W	1	25 W
Fan	10 W	1	10 W
System Total			213 W

Nameplate peak

Measured Peak 145 W
(Power-intensive workload)

In Google's world, for given DC power budget, deploy as many machines as possible

X. Fan, W-D Weber, L. Barroso, "Power Provisioning for a Warehouse-sized Computer," ISCA'07, San Diego, (June 2007).

Typical Datacenter Power

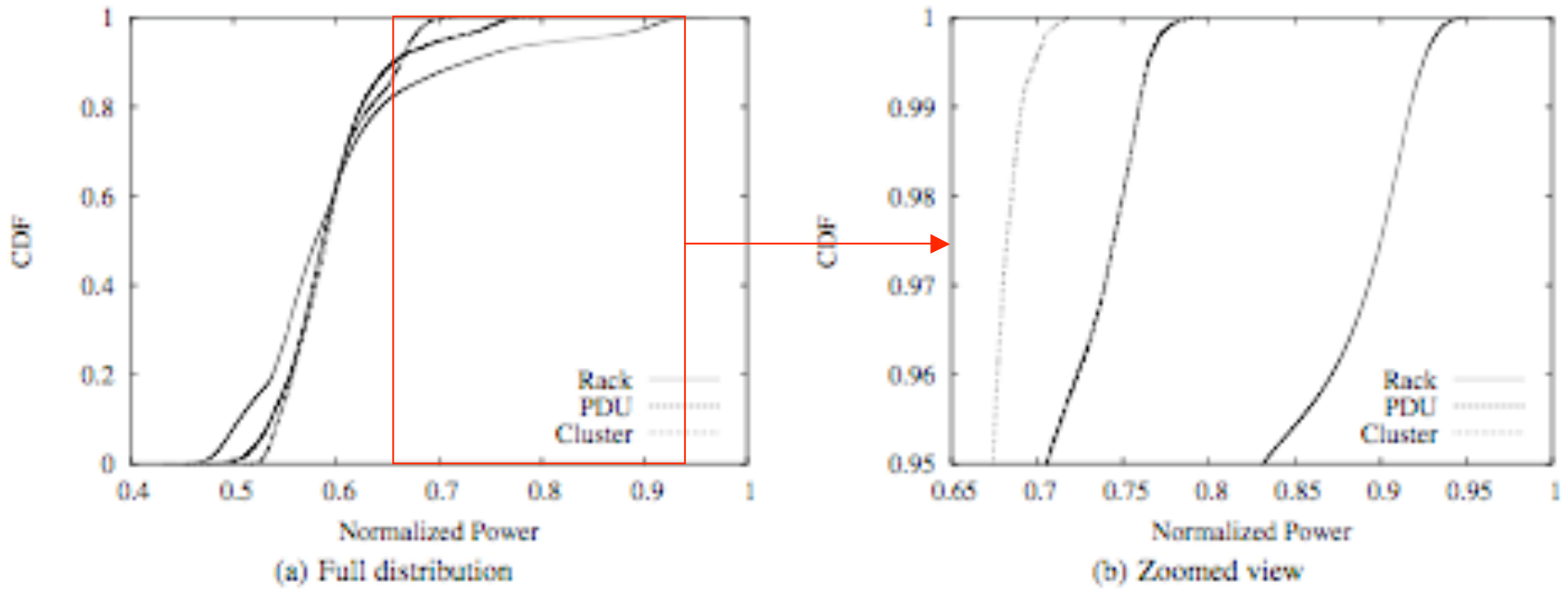
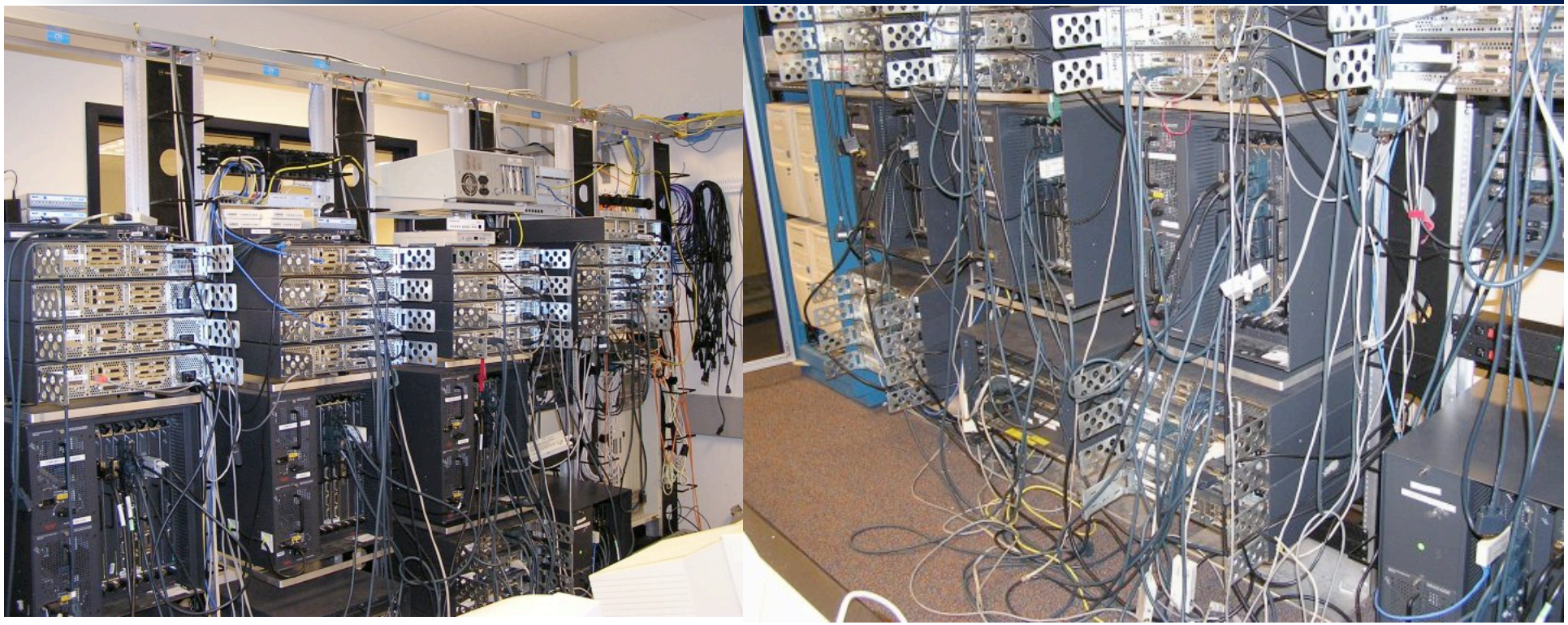


Figure 9: CDF of a Real Datacenter

Power-aware allocation of resources can achieve higher levels of utilization

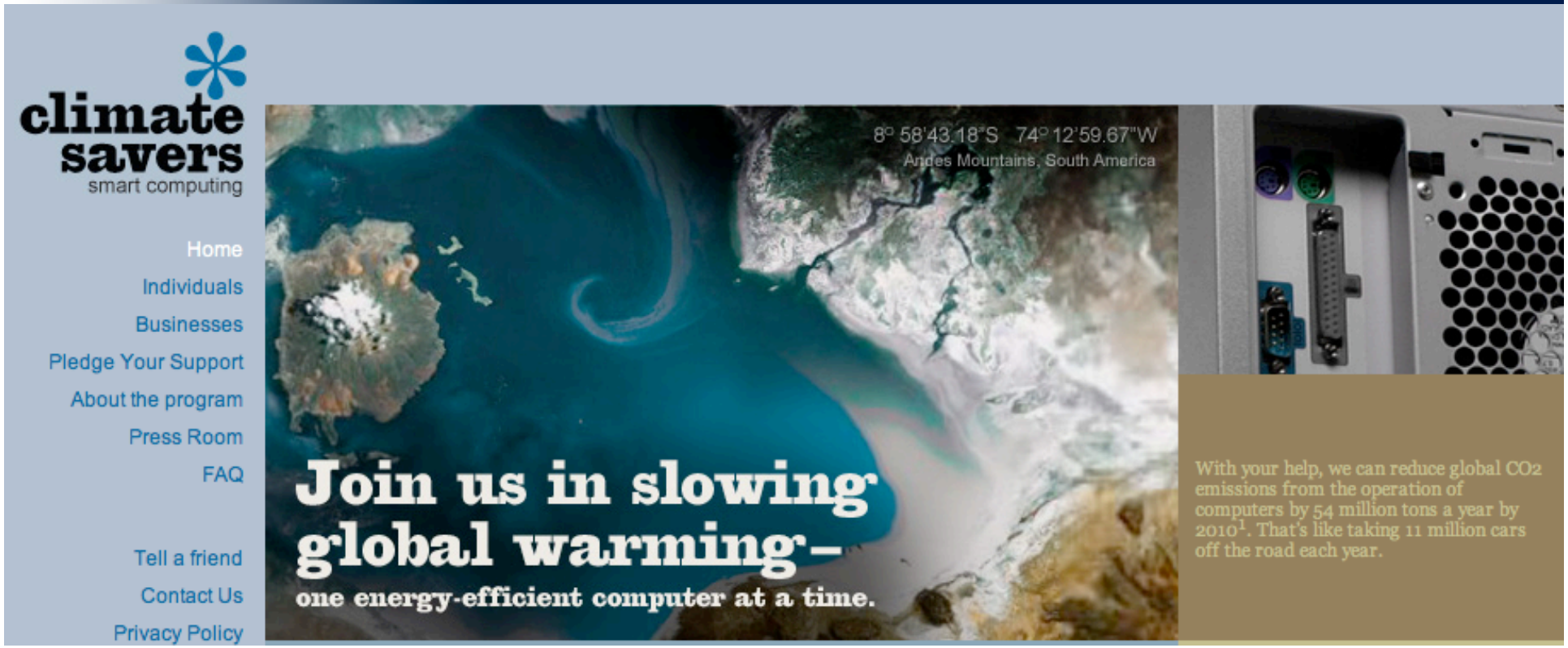
X. Fan, W-D Weber, L. Barroso, "Power Provisioning for a Warehouse-sized Computer," ISCA'07, San Diego, (June 2007).

FYI--Network Element Power



- 96 x 1 Gbit port Cisco datacenter switch consumes around 15 kW -- approximately 100x a typical dual processor Google server @ 145 W
- High port density drives network element design, but such high power density makes it difficult to tightly pack them with servers
- Is an alternative distributed processing/communications topology possible?

Climate Savers Initiative

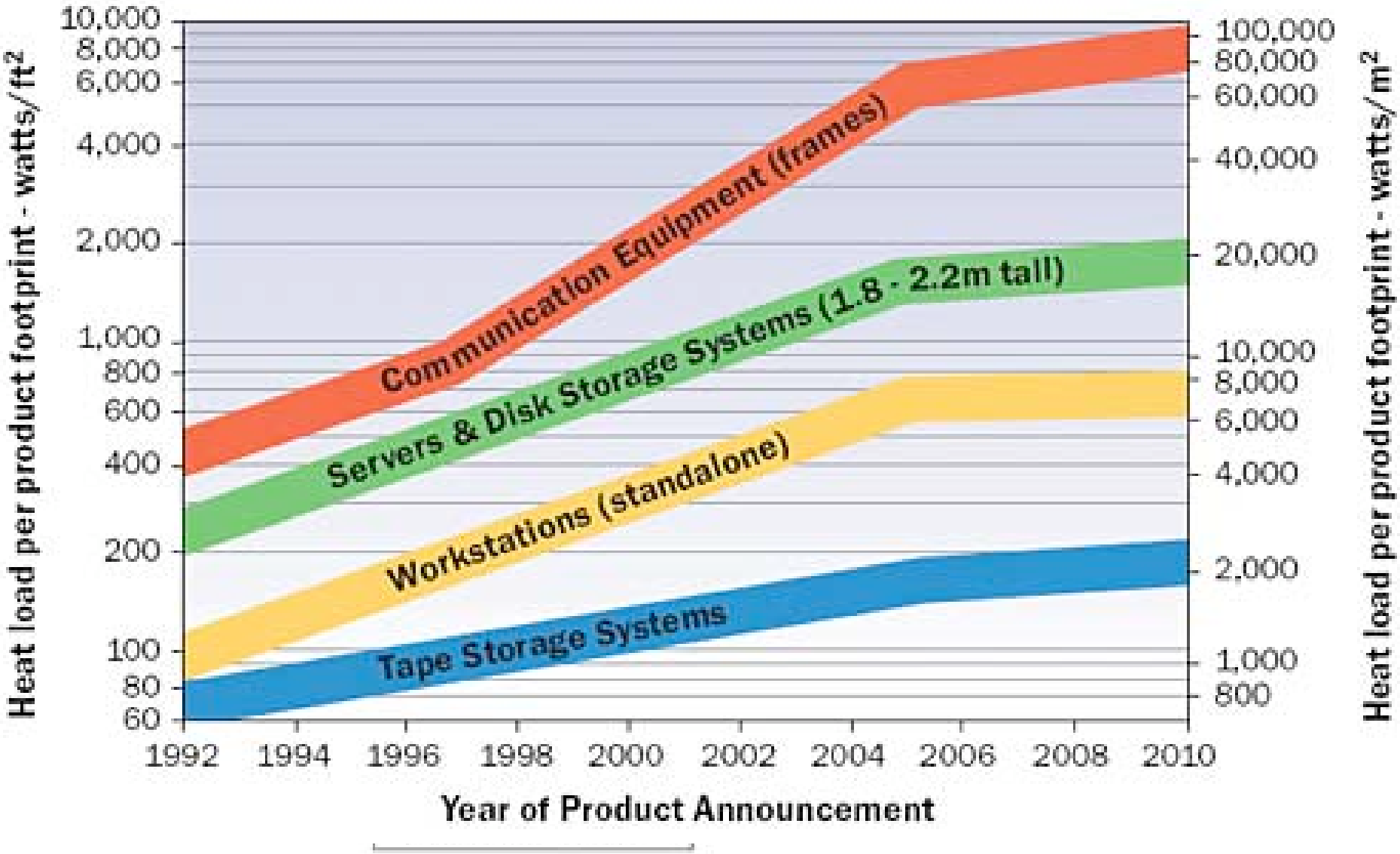


- Improving the efficiency of power delivery to computers as well as usage of power by computers
  - Transmission: 9% of energy is lost before it even gets to the datacenter
  - Distribution: 5-20% efficiency improvements possible using high voltage DC rather than low voltage AC
  - Chill air to mid 50s vs. low 70s to deal with the unpredictability of hot spots

DC Energy Conservation

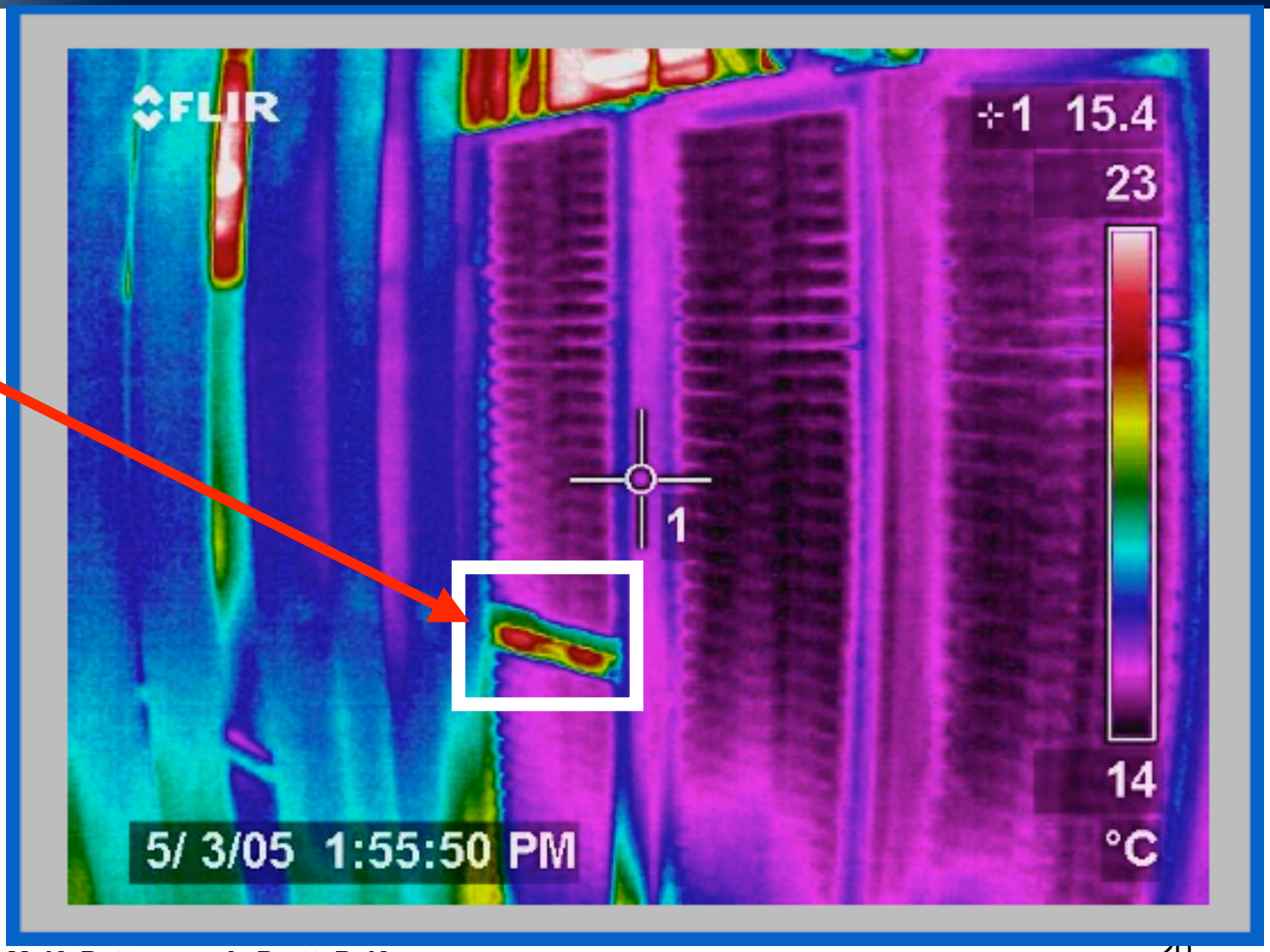
- DCs limited by power
  - For each dollar spent on servers, add $0.48 (2005)/$0.71 (2010) for power/cooling
  - $26B spent to power and cool servers in 2005 expected to grow to $45B in 2010
- Intelligent allocation of resources to applications
  - Load balance power demands across DC racks, PDUs, Clusters
  - Distinguish between user-driven apps that are processor intensive (search) or data intensive (mail) vs. backend batch-oriented (analytics)
  - Save power when peak resources are not needed by shutting down processors, storage, network elements

Power/Cooling Issues



Thermal Image of Typical Cluster Rack

Rack Switch



M. K. Patterson, A. Pratt, P. Kumar,
"From UPS to Silicon: an end-to-end evaluation of datacenter efficiency", Intel Corporation

DC Networking and Power

- Within DC racks, network equipment often the “hottest” components in the hot spot
- Network opportunities for power reduction
  - Transition to higher speed interconnects (10 Gbs) at DC scales and densities
  - High function/high power assists embedded in network element (e.g., TCAMs)

DC Networking and Power: Implications for QoS?

- Selectively power down ports/portions of net elements
- Enhanced power-awareness in the network stack
  - Power-aware routing and support for system virtualization
    - Support for datacenter “slice” power down and restart
  - Application and power-aware media access/control
    - Dynamic selection of full/half duplex
    - Directional asymmetry to save power, e.g., 10Gb/s send, 100Mb/s receive
  - Power-awareness in applications and protocols
    - Hard state (proxying), soft state (caching), protocol/data “streamlining” for power as well as b/w reduction
- Power implications for topology design
  - Tradeoffs in redundancy/high-availability vs. power consumption
  - VLANs support for power-aware system virtualization

Bringing Resources On-/Off-line

- Save power by taking DC “slices” off-line
  - Resource footprint of Internet applications hard to model
  - Dynamic environment, complex cost functions require measurement-driven decisions
  - Must maintain Service Level Agreements, no negative impacts on hardware reliability
  - Pervasive use of virtualization (VMs, VLANs, VStor) makes feasible rapid shutdown/migration/restart
- Recent results suggest that conserving energy may actually improve reliability
  - MTTF: stress of on/off cycle vs. benefits of off-hours

"System" Statistical Machine Learning

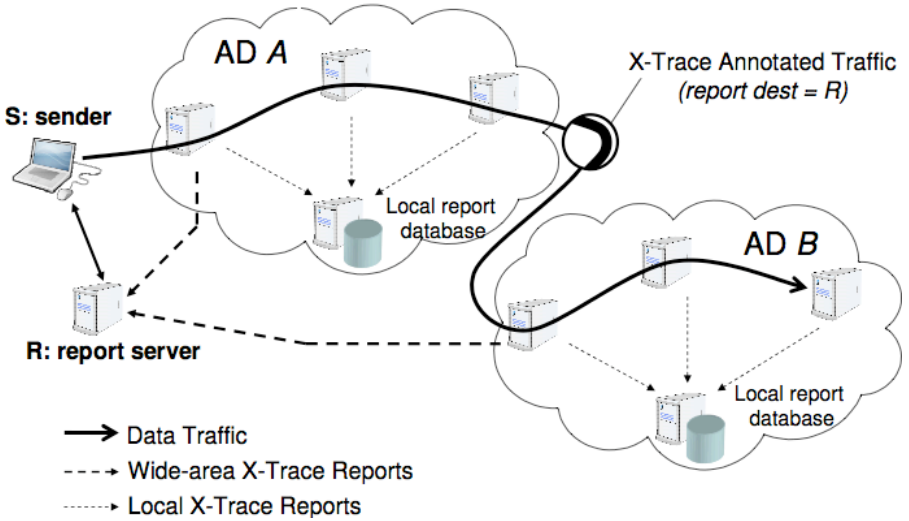
- S^2ML Strengths
  - Handle SW churn: Train vs. write the logic
  - Beyond queuing models: Learns how to handle/make policy between steady states
  - Beyond control theory: Coping with complex cost functions
  - Discovery: Finding trends, needles in data haystack
  - Exploit cheap processing advances: fast enough to run online
- S^2ML as an integral component of DC OS

Datacenter Monitoring

- To build models, S^2ML needs data to analyze -- the more the better!
- Huge technical challenge: trace 10K++ nodes within and between DCs
  - From applications across application tiers to enabling services
  - Across network layers and domains

RIOT: RadLab Integrated Observation via Tracing Framework

- Trace connectivity of distributed components
  - Capture **causal connections** between requests/responses
- Cross-layer
  - Include network and middleware services such as IP and LDAP
- Cross-domain
  - Multiple datacenters, composed services, overlays, mash-ups
  - Control to individual administrative domains



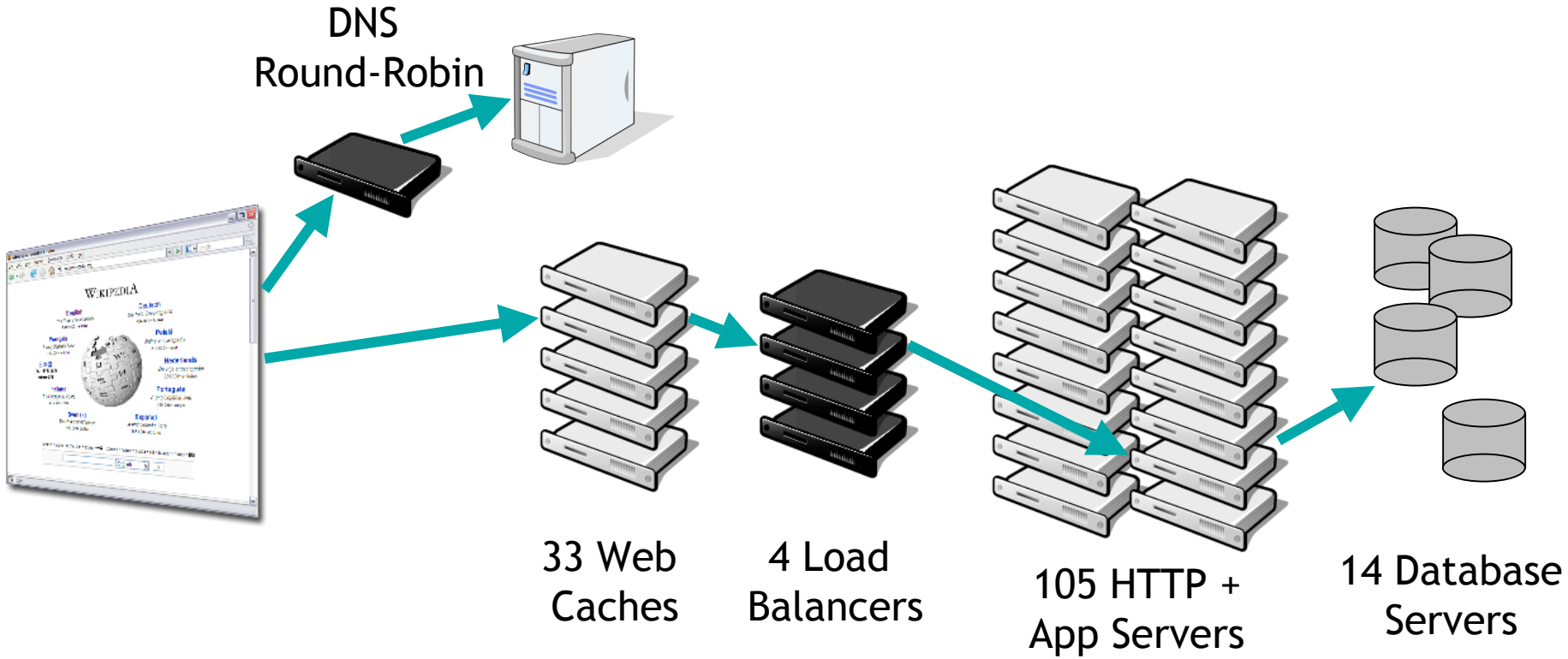
- "Network path" sensor
  - Put individual requests/responses, at different network layers, in the context of an end-to-end request

X-Trace: Path-based Tracing

- Simple and universal framework
  - Building on previous path-based tools
  - Ultimately, every protocol and network element should support tracing
- Goal: end-to-end path traces with today's technology
  - *Across the whole network stack*
  - Integrates *different applications*
  - Respects *Administrative Domains*' policies

Example: Wikipedia

Many servers, four worldwide sites

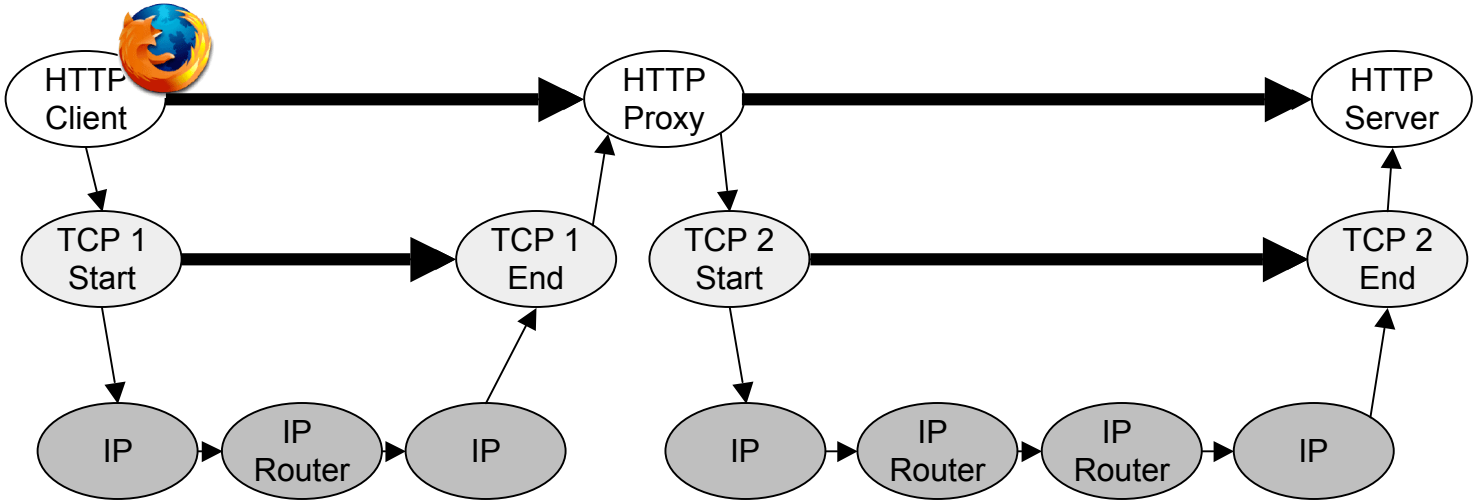


A user gets a stale page: What went wrong?

Four levels of caches, network partition, misconfiguration, …

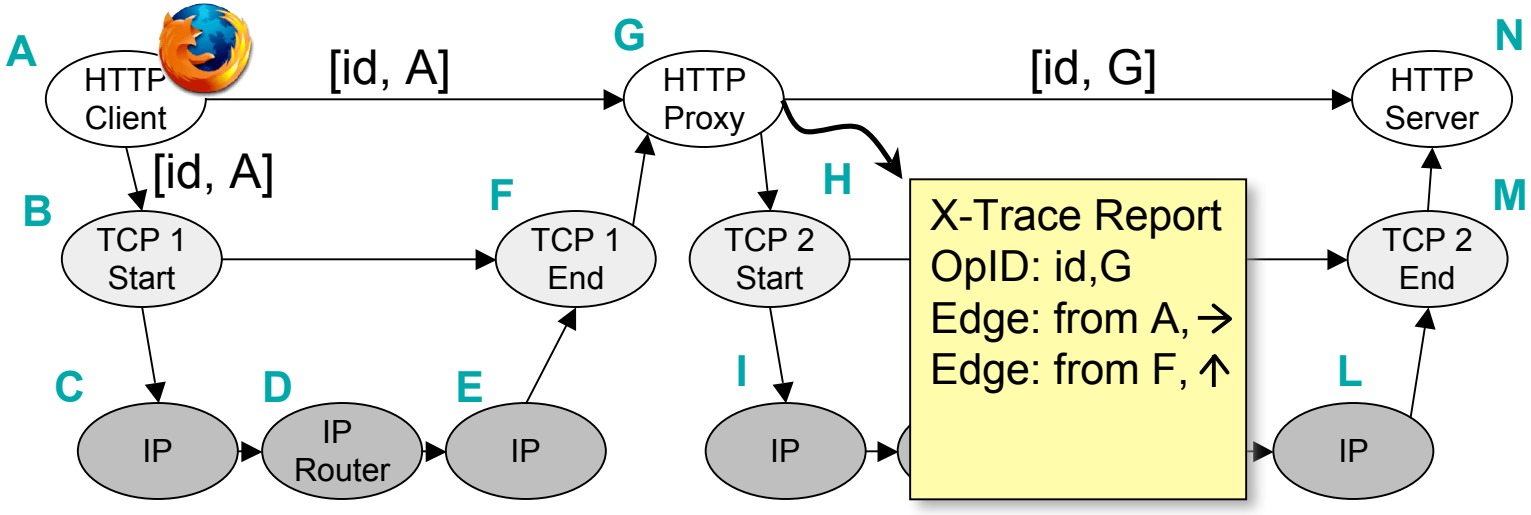
Task

- Specific system activity in the datapath
  - E.g., sending a message, fetching a file
- Composed of many *operations (or events)*
  - Different abstraction levels
  - Multiple layers, components, domains



Task graphs can be named, stored, and analyzed

Basic Mechanism

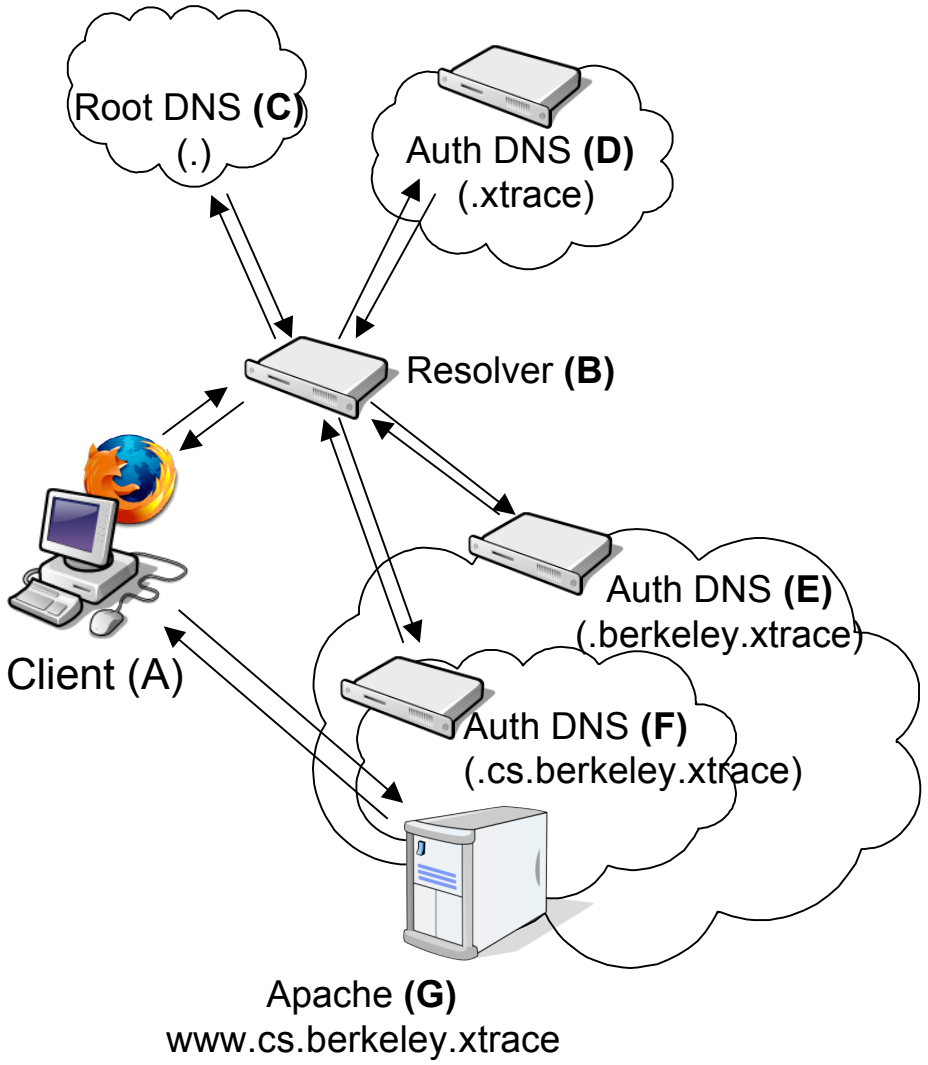


- All operations: same ***TaskId***
- Each operation: unique ***OpId***
- ***Propagate*** on edges: [TaskId, OpId]
- Nodes ***report*** all incoming edges to a collection infrastructure

Some Details

- Metadata
  - Horizontal edges: encoded in protocol messages
    - (HTTP headers, TCP options, etc.)
  - Vertical edges: widened API calls (setsockopt, etc.)
- Each device
  - Extracts the metadata from incoming edges
  - Creates an new operation ID
  - Copies the updated metadata into outgoing edges
  - Issues a report with the new operation ID
    - Set of key-value pairs with information about operation
- No Layering Violation
  - Propagation among same and adjacent layers
  - Reports contain information about specific layer

Example: DNS + HTTP

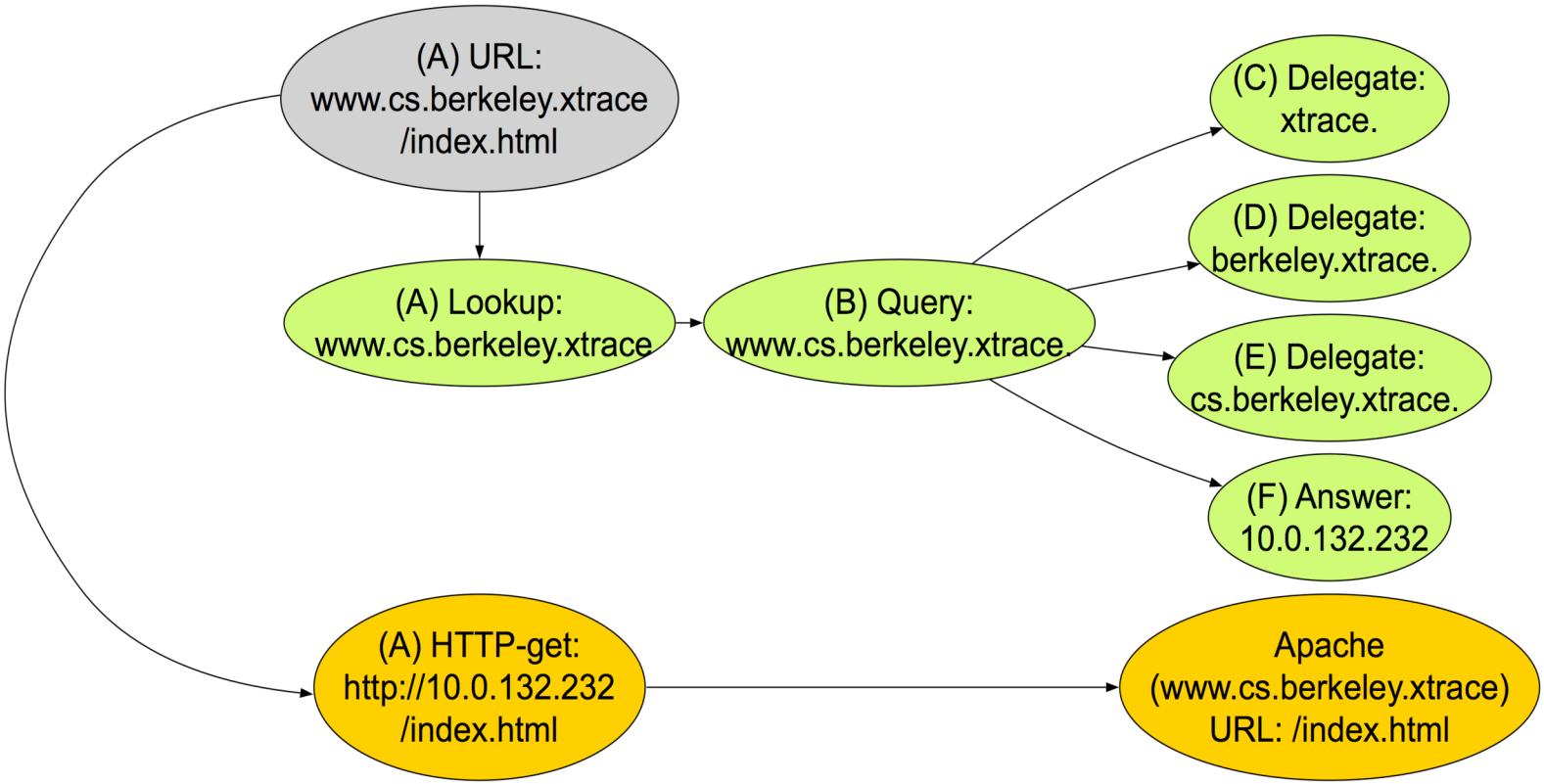


- Different applications
- Different protocols
- Different Administrative domains

- (A) through (F) represent 32-bit random operation IDs

Example: DNS + HTTP

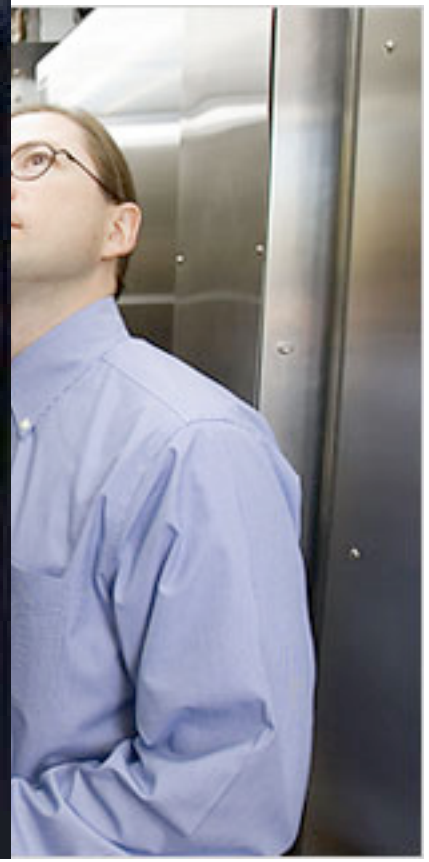
- Resulting X-Trace Task Graph



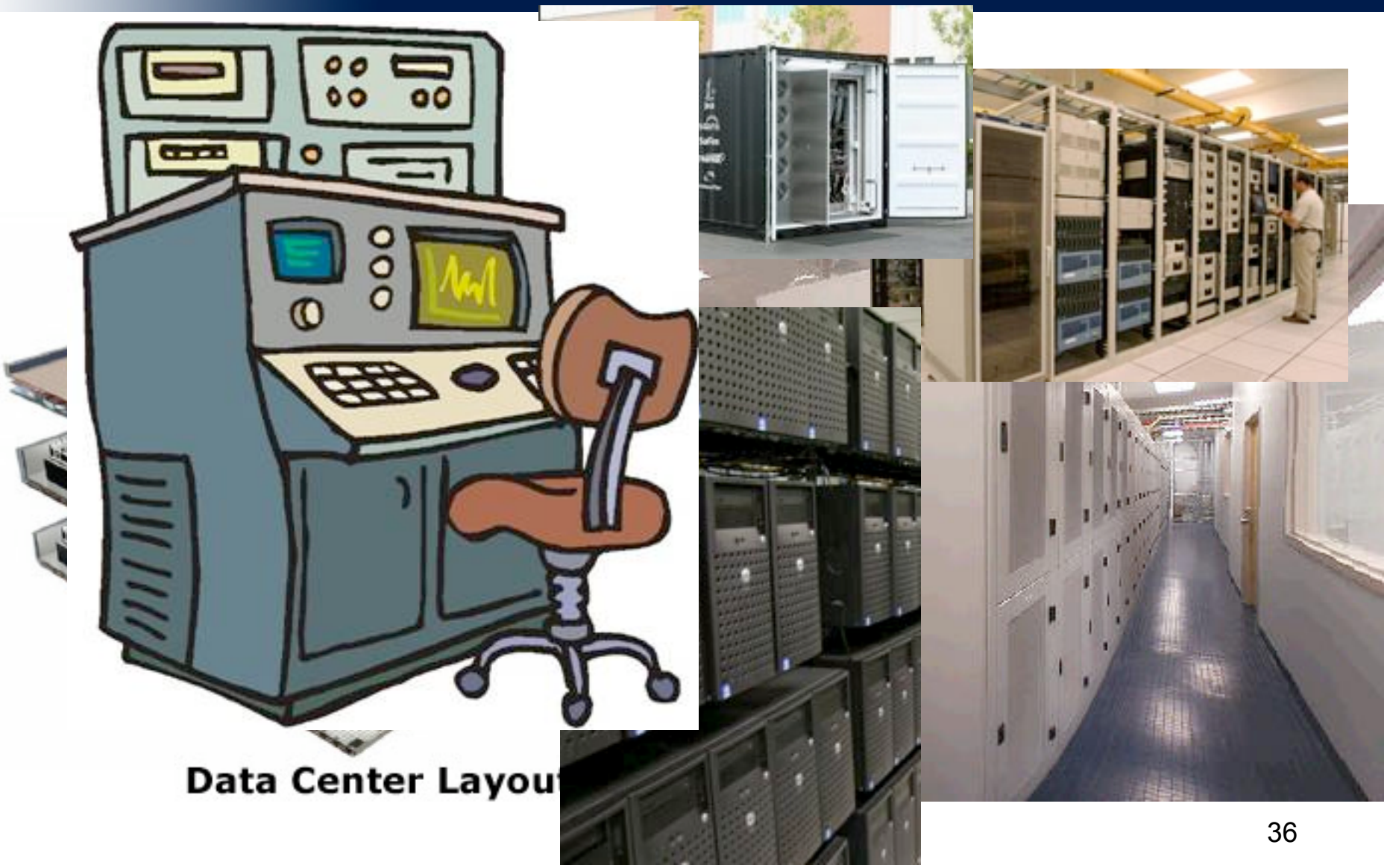
Summary and Conclusions

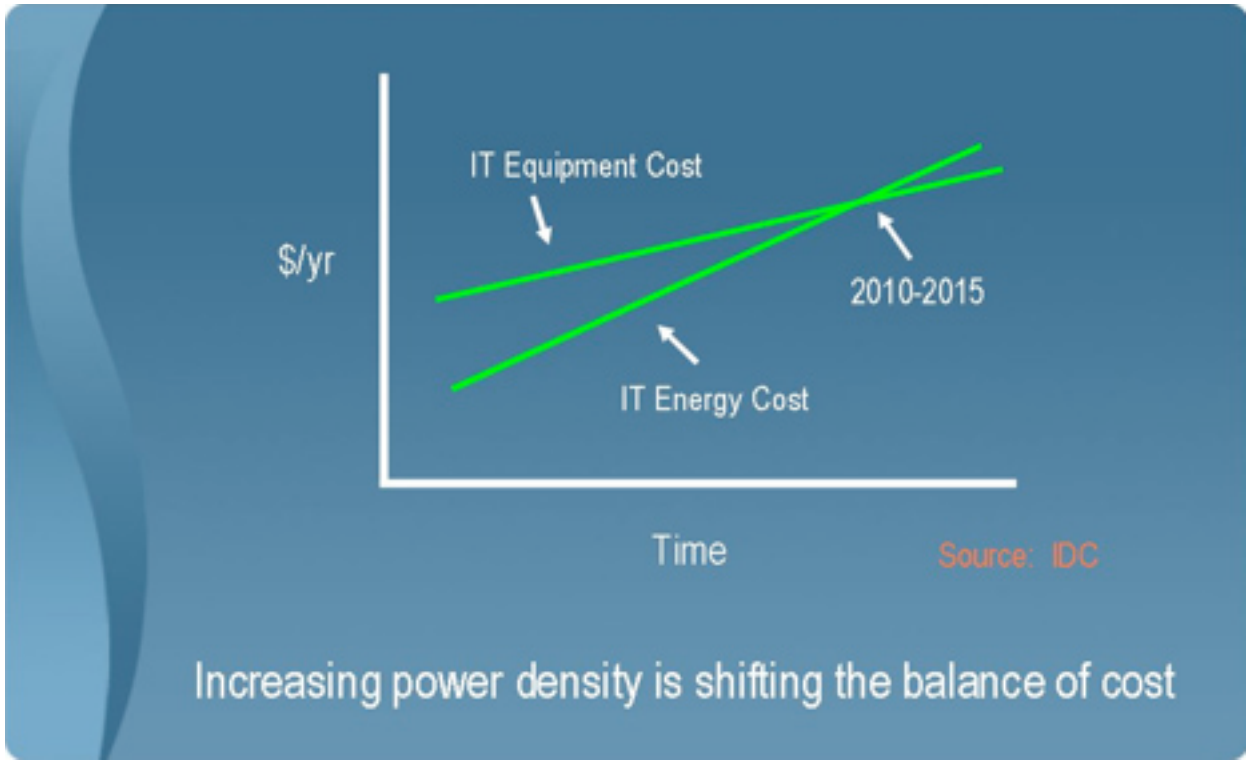
- Where the QoS Action is: Internet Datacenters
  - It's the backend to billions of network capable devices
  - Plenty of heap processing, storage, and bandwidth
  - Challenge: use it efficiently from an energy perspective
- DC Network Power Efficiency is an Emerging Problem!
  - Much known about processors, little about networks
  - Faster/denser network fabrics stressing power limits
- Enhancing Energy Efficiency and Reliability
  - Must consider the whole stack from client to web application running in the datacenter
  - Power- and network-aware resource allocation
  - Trade latency/thruput for power by shutting down resources
  - Predict workload patterns to bring resources on-line to satisfy SLAs, particularly user-driven/latency-sensitive applications
  - Path tracing + SML to uncover correlated behavior of network and application services

Thank You!



Internet Datacenter




IT Equipment Cost
$/yr
2010-2015
IT Energy Cost
Time
Source: IDC
Increasing power density is shifting the balance of cost