ΕΠΛ605 Προχωρημένη Αρχιτεκτονική Υπολογιστών

Data Center Scale Architecture

Data Center and TCO Analysis

Edge vs Cloud Datacenter

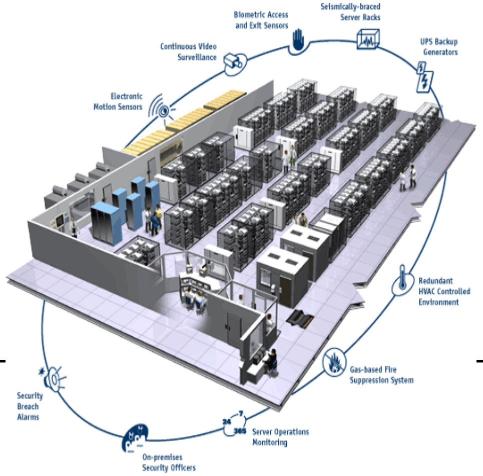
Παναγιώτα Νικολάου

Χειμερινό Εξάμηνο 2018

Slides based on:

1. Hennesy and Patterson CAQA 5th edition Morgan and Kaufman
2. Anand Sivasubramaniam Dept. of Computer Science & Eng. The Pennsylvania State University
3. Barroso, Luiz André, Jimmy Clidaras, and Urs Hölzle. "The datacenter as a computer: An introduction to the design of warehouse-scale machines."

Datacenters: The What?



- Large numbers of servers, storage devices, and network switches housed in a single facility

- Types of Datacenters:
  - Traditional Enterprise Datacenters
  - Warehouse Scale Computers

Kinds of Datacenters

Traditional DCs

- Consolidated infrastructure of different organizational units
- Large number of small/medium sized apps
- Often 3rd party software
- E.g. Those in Financial Services, Pharma, etc.

Warehouse-Scale Computers

- Single organization/unit
- Small number of very large apps
- Software built in-house
- E.g. Google, Microsoft, Facebook, etc.

Introduction

- Warehouse-scale computer (WSC)
  - Provides Internet services
    - Search, social networking, online maps, video sharing, online shopping, email, cloud computing, etc.
  - Differences with HPC “clusters”:
    - Clusters have higher performance processors and network
    - Clusters emphasize thread-level parallelism, WSCs emphasize request-level parallelism

Introduction

- Important design factors for WSC:
  - Cost-performance
    - Small savings add up
  - Energy efficiency
    - Affects power distribution and cooling
    - Work per joule
  - Dependability via redundancy
    - Commodity HW and software based redundancy
  - Network I/O
    - Consistency and interface to world
  - Interactive and batch processing workloads
    - Search but also calculate meta data (rank pages)
    - Online and off-line jobs, collocation

Introduction

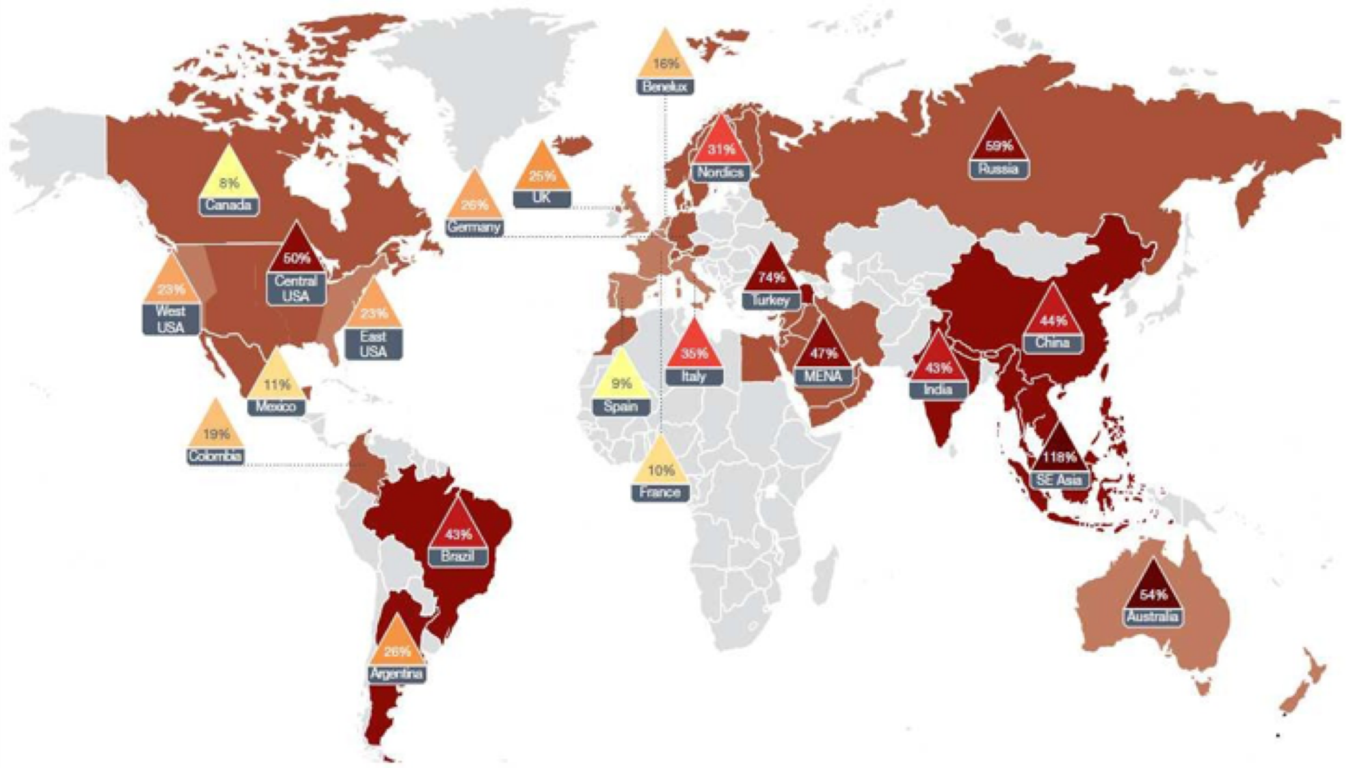
- Important design factors for WSC:
  - Computational parallelism is not important
    - Most jobs are totally independent
    - "Request-level parallelism"
    - Mostly read and often write to not shared data
    - Data in storage for batch
  - Operational costs count
    - Power consumption is a primary, not secondary, constraint when designing system
  - Scale and its opportunities and problems
    - Can afford to build customized systems since WSC require volume purchase.
    - Large scale means failures more common

Some statistics

- Number of Datacenters: 510,000 [Src: Emerson, 2011]
- Real Estate of Datacenters: 285 Million $(~6000 football fields) [Src: Emerson, 2011]
- New Investments in 2012: $105 Billion, up from $ 86 Billion in 2011 [Src: Computer Weekly]
- Power Demand in 2012: 38 GW, up from 24 GWin 2011 [Src: Computer Weekly]
- Electricity Consumption in 2010: 285 Billion KWH, 1.5% of world consumption [Src: NY Times]. If IT Sector viewed as a country, fifth largest electricity consumer.

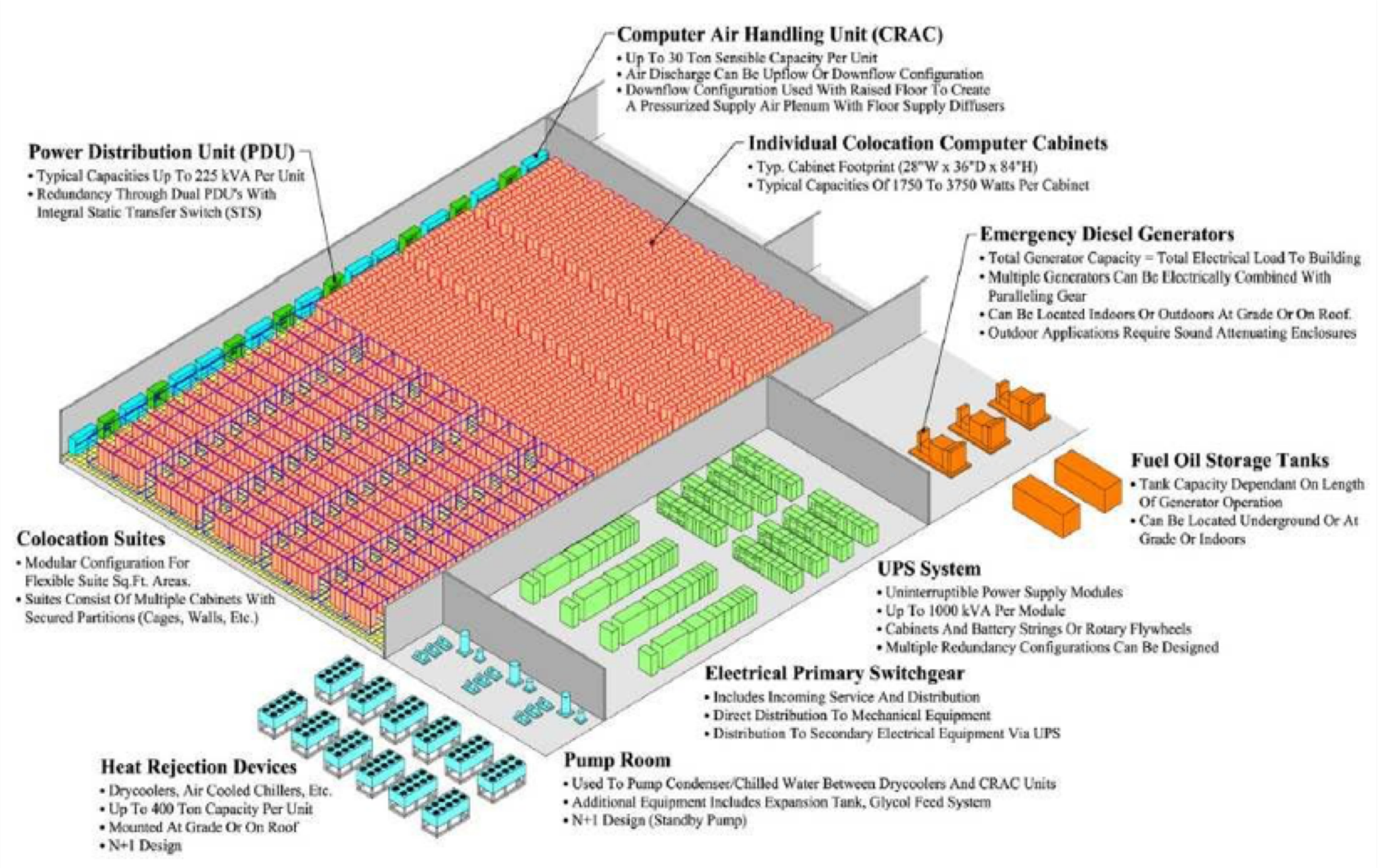
- Carbon Footprint: 2% of Global footprint

Continuing Growth in Datacenters



Projected Growth 2011-2012. Src: Datacenter Dynamics

Datacenter Infrastructure



Computer Air Handling Unit (CRAC)
- Up To 30 Ton Sensible Capacity Per Unit
- Air Discharge Can Be Upflow Or Downflow Configuration
- Downflow Configuration Used With Raised Floor To Create A Pressurized Supply Air Plenum With Floor Supply Diffusers

Power Distribution Unit (PDU)
- Typical Capacities Up To 225 kVA Per Unit
- Redundancy Through Dual PDU's With Integral Static Transfer Switch (STS)

Individual Colocation Computer Cabinets
- Typ. Cabinet Footprint (28"W x 36"D x 84"H)
- Typical Capacities Of 1750 To 3750 Watts Per Cabinet

Emergency Diesel Generators
- Total Generator Capacity = Total Electrical Load To Building
- Multiple Generators Can Be Electrically Combined With Paralleling Gear
- Can Be Located Indoors Or Outdoors At Grade Or On Roof.
- Outdoor Applications Require Sound Attenuating Enclosures

Fuel Oil Storage Tanks
- Tank Capacity Dependant On Length Of Generator Operation
- Can Be Located Underground Or At Grade Or Indoors

Colocation Suites
- Modular Configuration For Flexible Suite Sq.Ft. Areas.
- Suites Consist Of Multiple Cabinets With Secured Partitions (Cages, Walls, Etc.)

UPS System
- Uninterruptible Power Supply Modules
- Up To 1000 kVA Per Module
- Cabinets And Battery Strings Or Rotary Flywheels
- Multiple Redundancy Configurations Can Be Designed

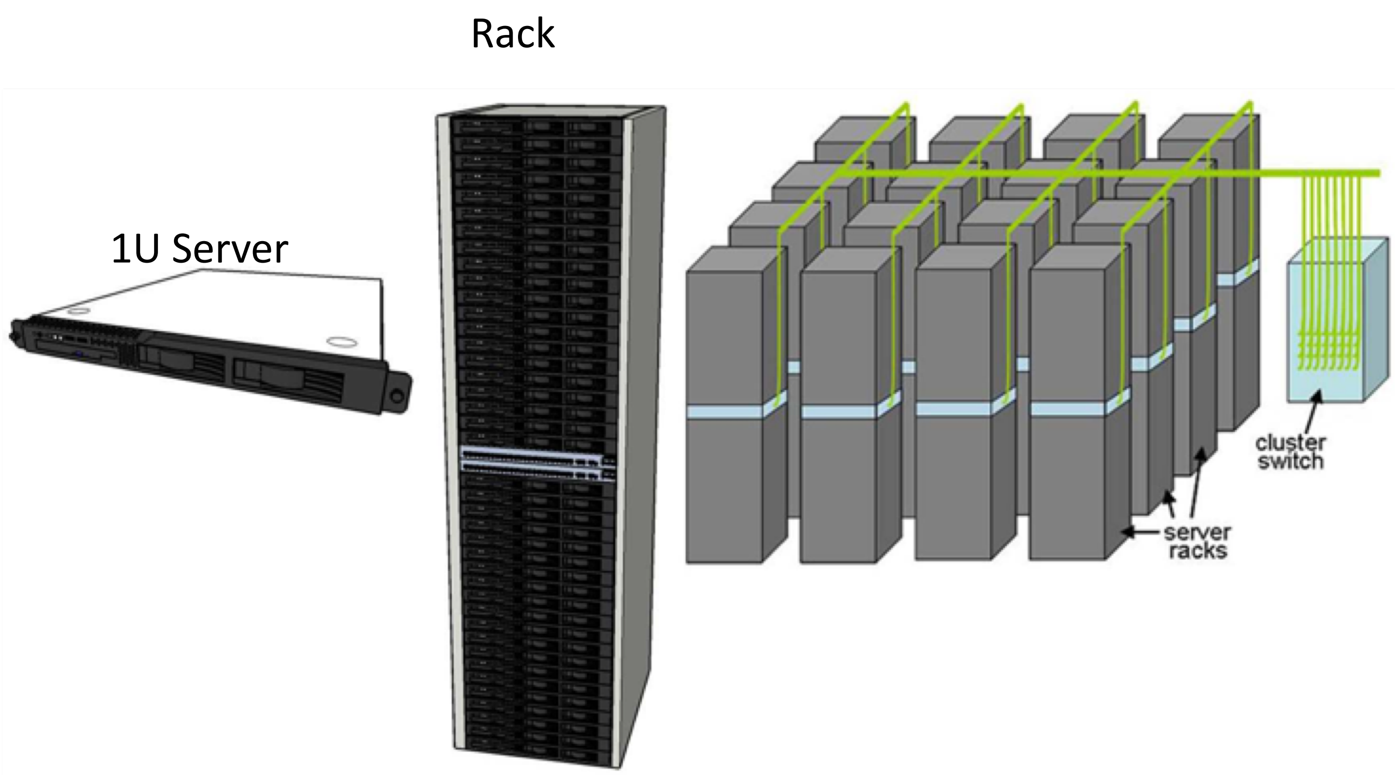
Electrical Primary Switchgear
- Includes Incoming Service And Distribution
- Direct Distribution To Mechanical Equipment
- Distribution To Secondary Electrical Equipment Via UPS

Heat Rejection Devices
- Drycoolers, Air Cooled Chillers, Etc.
- Up To 400 Ton Capacity Per Unit
- Mounted At Grade Or On Roof
- N+1 Design

Pump Room
- Used To Pump Condenser/Chilled Water Between Drycoolers And CRAC Units
- Additional Equipment Includes Expansion Tank, Glycol Feed System
- N+1 Design (Standby Pump)

Computer Architecture of WSC

The Computing Infrastructure



Computer Architecture of WSC

- WSC often use a hierarchy of networks for interconnection
- Each 19” rack holds 48 1U servers connected to a rack switch
- Rack switches are uplinked to switch higher in hierarchy
  - Uplink has 48 / n times lower bandwidth, where n = # of uplink ports
    - “Oversubscription”
  - Goal is to maximize locality of communication relative to the rack

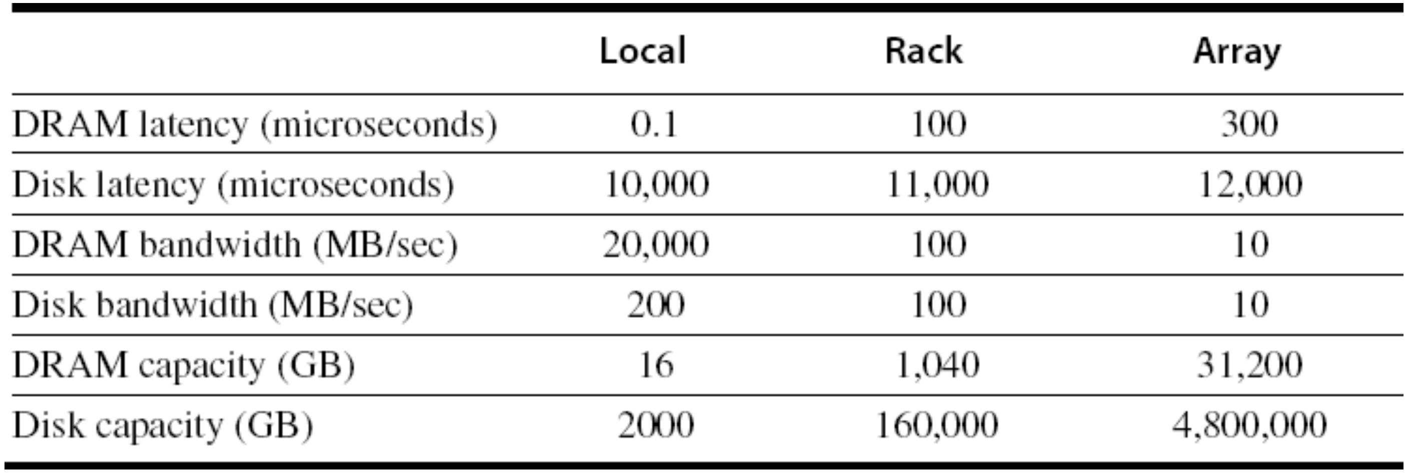
Storage

- Storage options:
  - Use disks inside the servers, or
  - Network attached storage through Infiniband

  - WSCs generally rely on local disks
  - Google File System (GFS) uses local disks and maintains at least three replicas
    - Magnetic storage vs SSD

WSC Memory Hierarchy

- **Servers can access DRAM and disks on other servers using a NUMA-style interface**



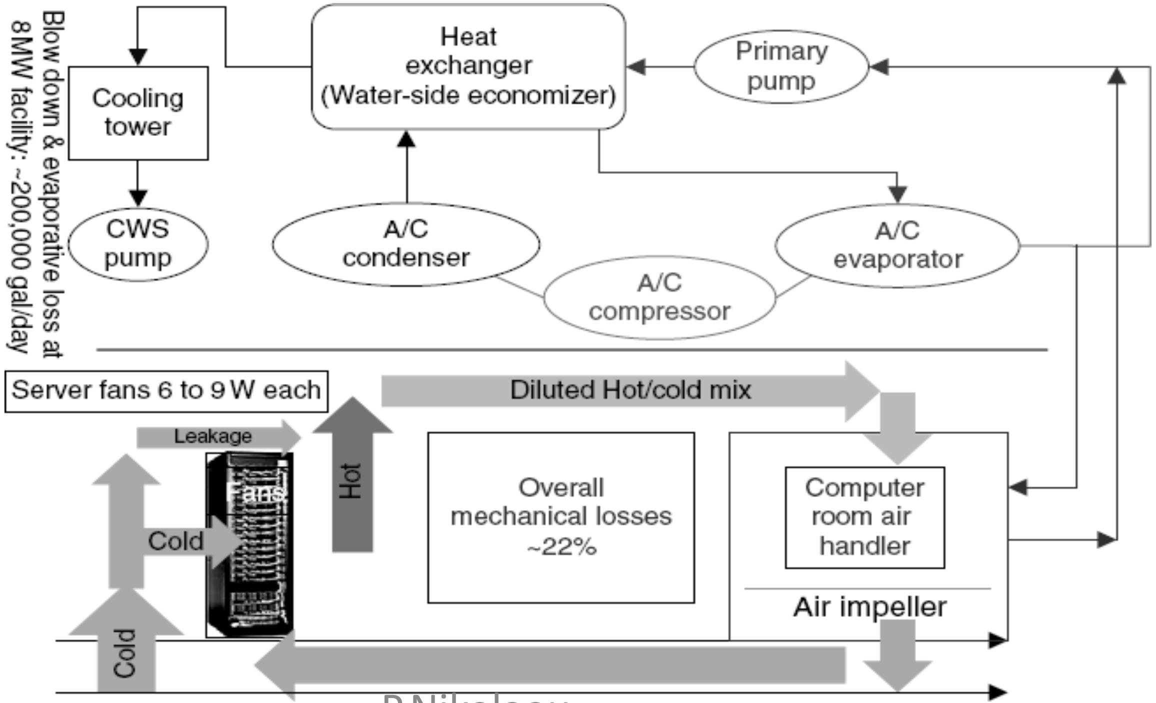
	Local	Rack	Array
DRAM latency (microseconds)	0.1	100	300
Disk latency (microseconds)	10,000	11,000	12,000
DRAM bandwidth (MB/sec)	20,000	100	10
Disk bandwidth (MB/sec)	200	100	10
DRAM capacity (GB)	16	1,040	31,200
Disk capacity (GB)	2000	160,000	4,800,000

Infrastructure and Cost correlation of WSC (CAPEX cost)

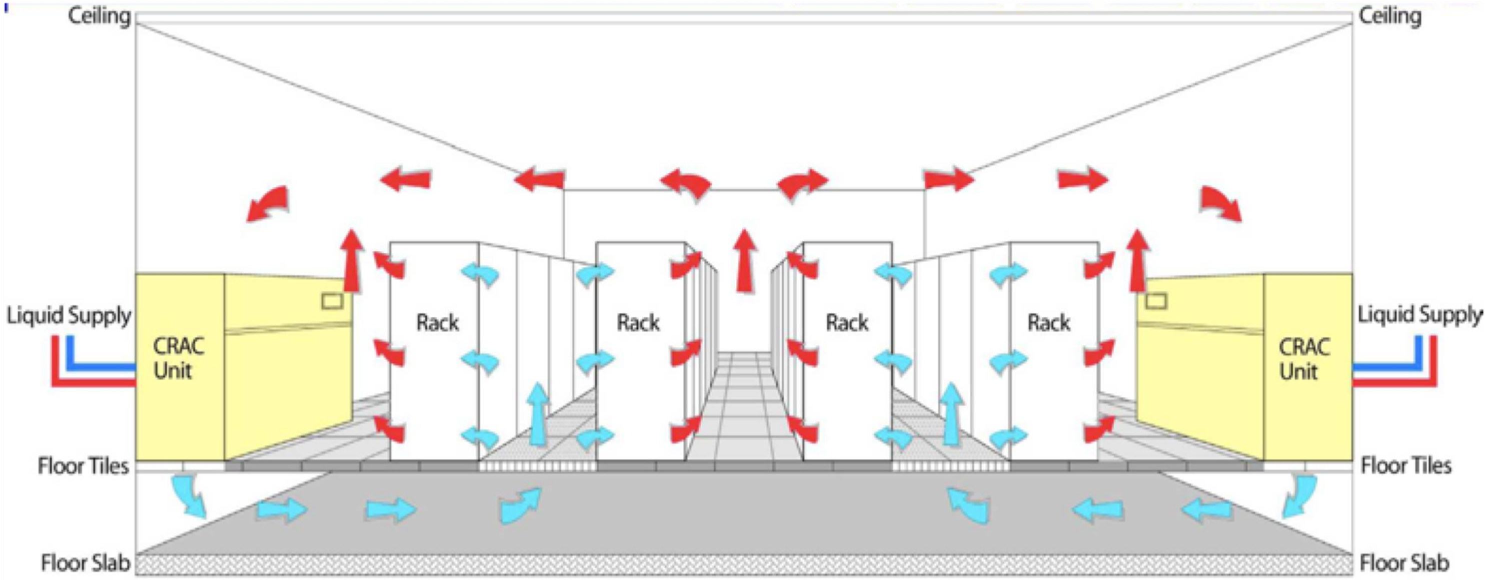
- Location of WSC
  - Proximity to Internet backbones, electricity cost, property tax rates, low risk from earthquakes, floods, and hurricanes

Infrastructure and Costs of WSC

- **Cooling**
  - **Air conditioning used to cool server room**
  - **64 F – 71 F**
    - **Keep temperature higher (closer to 71 F)**
  - **Cooling towers can also be used**
    - **Minimum temperature is “wet bulb temperature”**


Blow down & evaporative loss at 8MW facility: ~200,000 gal/day
Cooling tower
CWS pump
Heat exchanger (Water-side economizer)
Primary pump
A/C condenser
A/C compressor
A/C evaporator
Server fans 6 to 9 W each
Diluted Hot/cold mix
Leakage
Fans
Hot
Cold
Cold
Overall mechanical losses ~22%
Computer room air handler
Air impeller

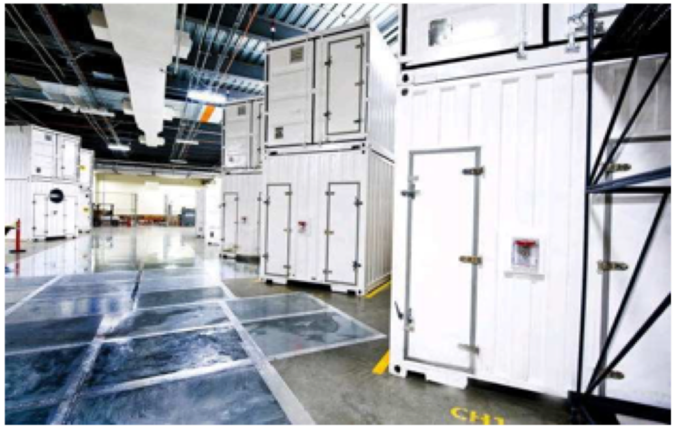
Cooling: Inside the Machine Room



- CRAC = Computer Room Air Conditioning
- Cold Air goes into servers (sucked in by fans) from Cold Aisles and comes out to the Hot Aisle
- Cold Aisles ~18-22C, Hot Aisles > 35C

Container-Based Datacenters

Microsoft DC in Chicago, IL



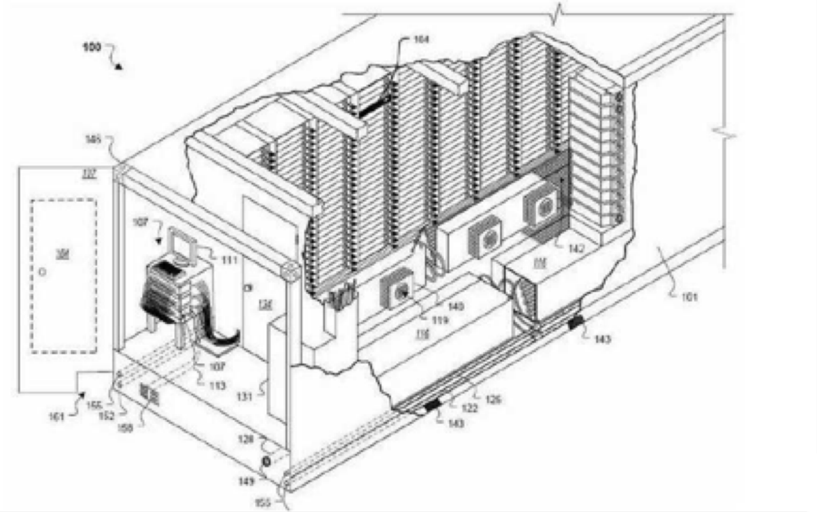
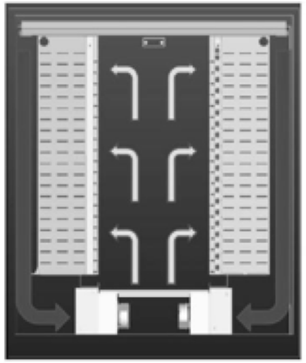


Figure 6.19 Google customizes a standard 1AAA container: 40 x 8 x 9.5 feet (12.2 x 2.4 x 2.9 meters). The servers are stacked up to 20 high in racks that form two long rows of 29 racks each, with one row on each side of the container. The cool aisle goes down the middle of the container, with the hot air return being on the outside. The hanging rack structure makes it easier to repair the cooling system without removing the servers. To allow people inside the container to repair components, it contains safety systems for fire detection and mist-based suppression, emergency egress and lighting, and emergency power shut off. Containers also have many sensors: temperature, airflow pressure, air leak detection, and motion-sensing lighting. A video tour of the datacenter is found at http://www.google.com/corporate/green/datacenters/summit.html. Microsoft, Yahoo, and many others are now building modular datacenters based upon these ideas but they have stopped using ISO standard containers since the size is inconvenient.

Google, Circa 2007

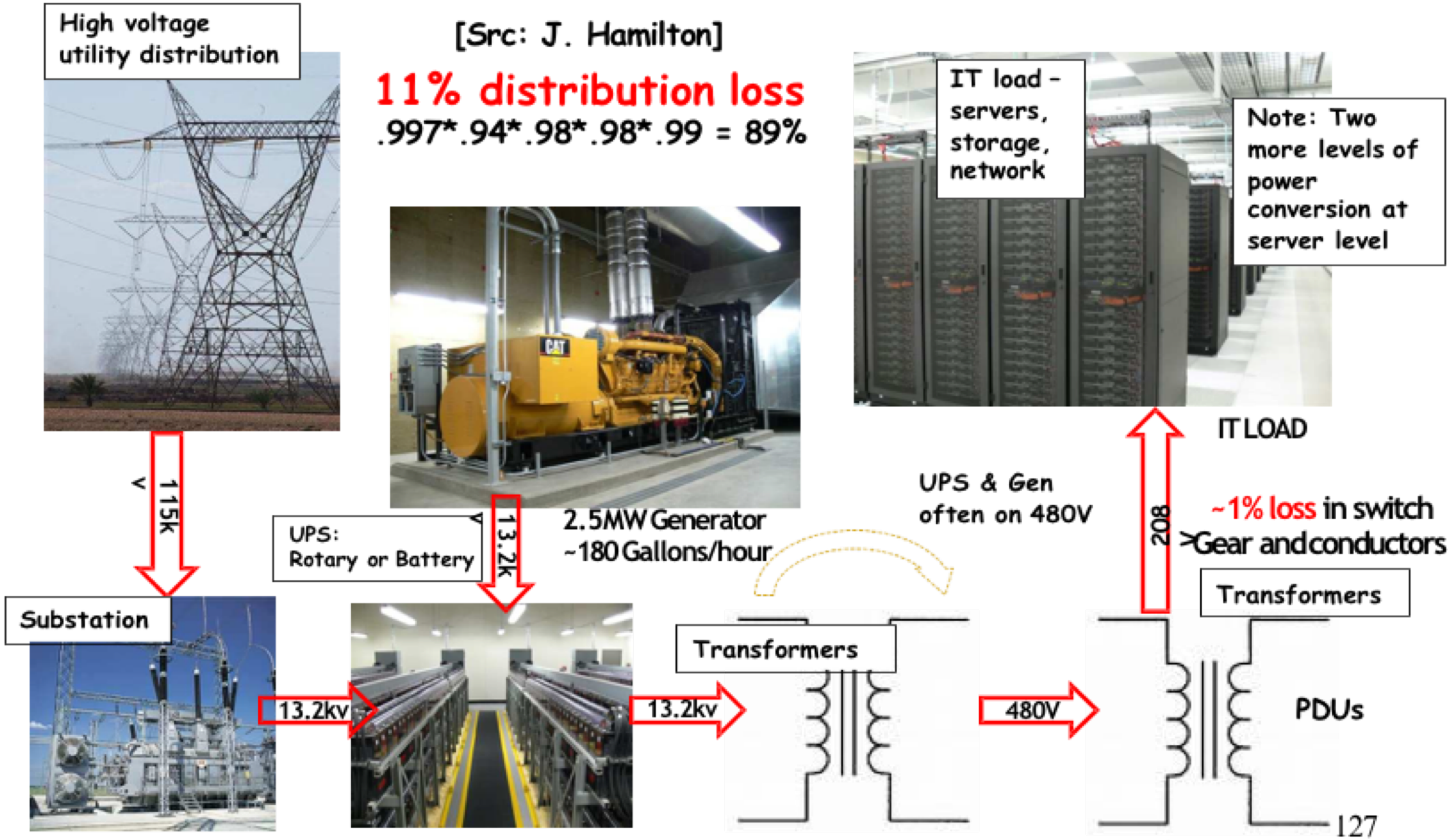


Power Infrastructure

[Src: J. Hamilton]

11% distribution loss

.997*.94*.98*.98*.99 = 89%



High voltage utility distribution

115kv

Substation

99.7% efficient

13.2kv

UPS: Rotary or Battery

13.2kv

2.5MW Generator
~180 Gallons/hour

94% efficient

13.2kv

Transformers

98% efficient

UPS & Gen often on 480V

480V

Transformers

PDUs

208

127

98% efficient

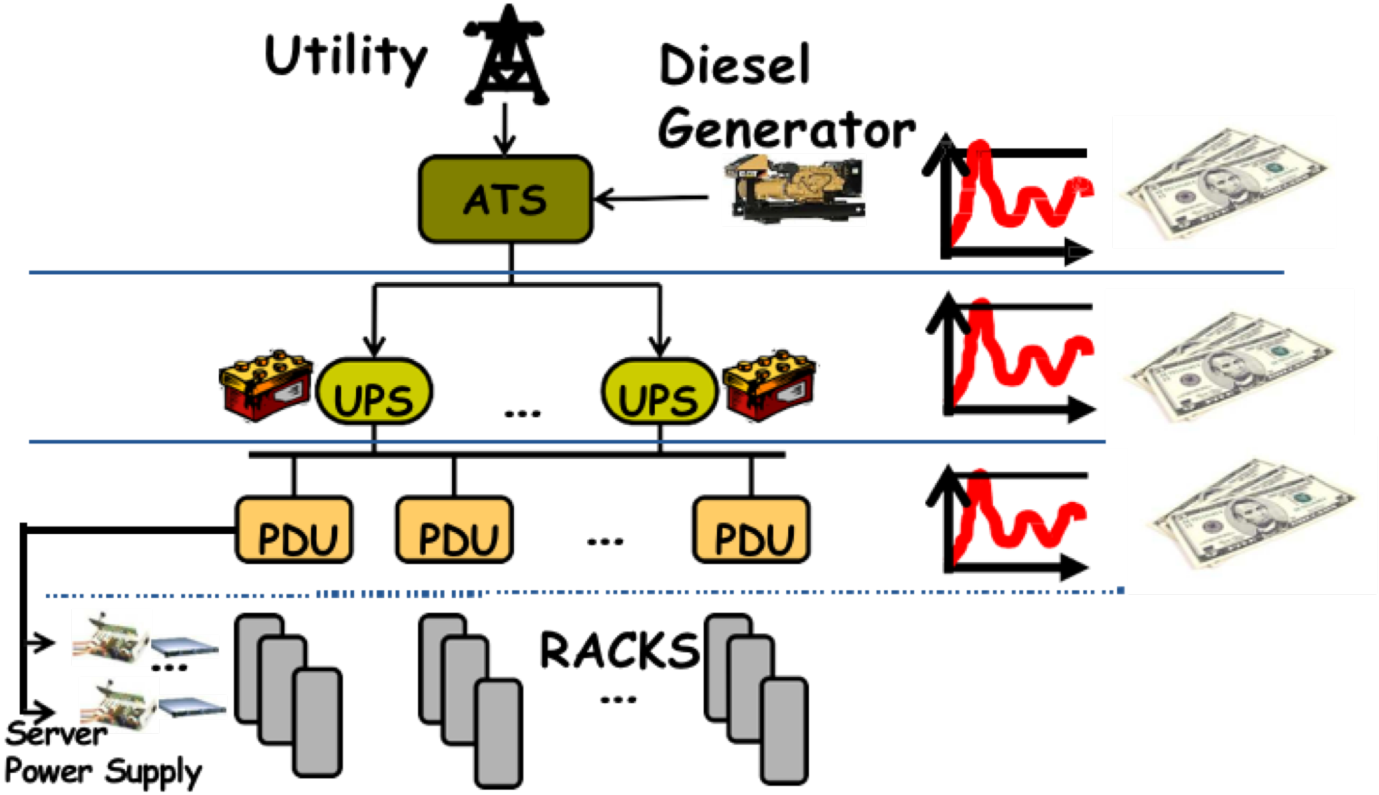
~1% loss in switch
Gear and conductors

IT LOAD

IT load - servers, storage, network

Note: Two more levels of power conversion at server level

Peak Power – Cap-Ex Costs



Costs further shoot up with the need for redundancy!

Datacenter Opex costs

- Maintenance
- Energy
- Personnel Costs

- **Maintenance:**
  - **Use replicas of data across different servers**
  - **If one slow or fails start on another**
  - **Use relaxed consistency:**
    - **No need for all replicas to always agree**

Causes of Outages and Anomalies (2400 servers 1st yr)

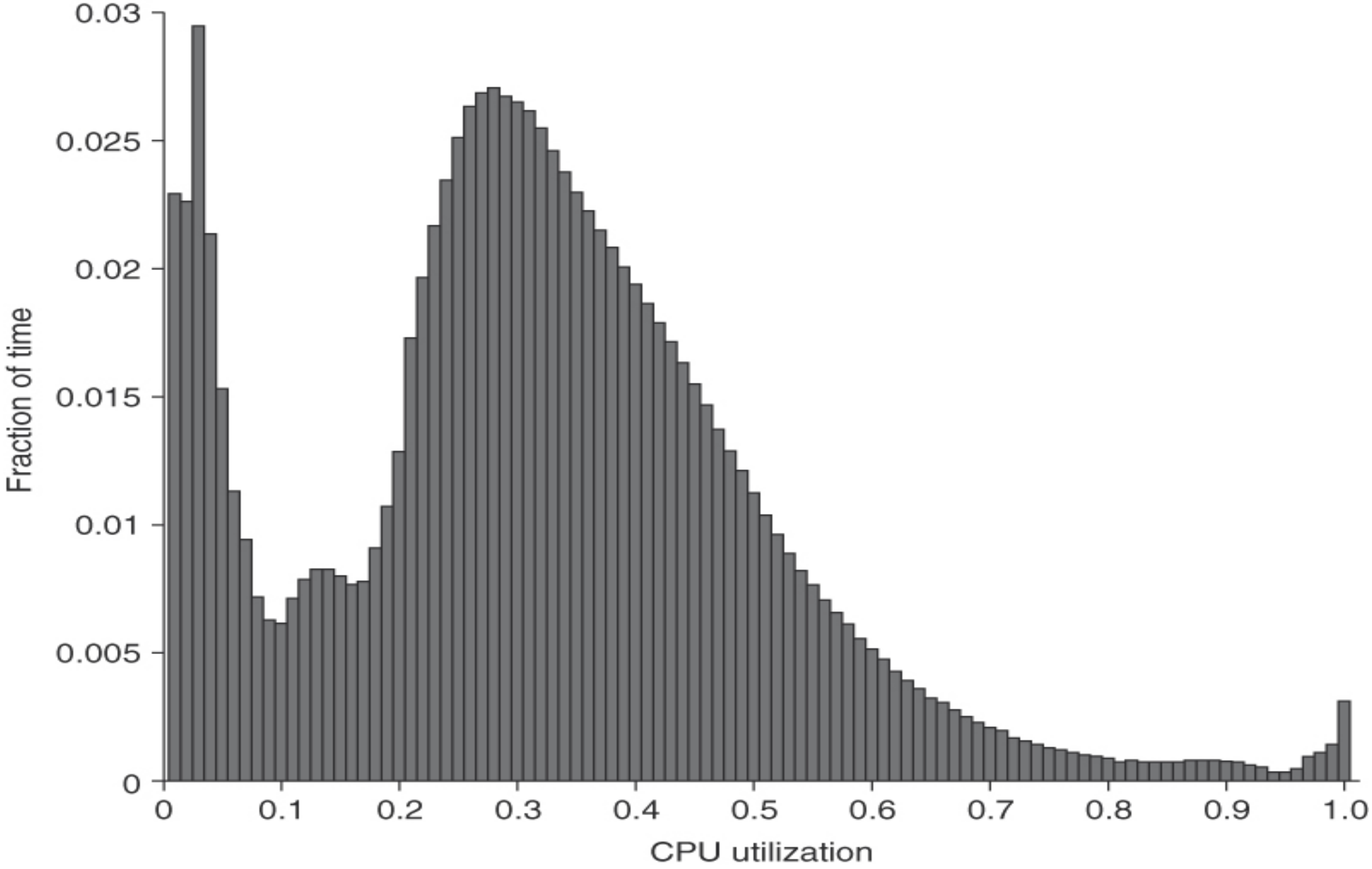
Number	Cause
1-2	Power outage
4	Upgrades
1000s	Hard-disk
	Dram
	Problematic machines
5000	Server crashes

Difference between server and service unavailability.

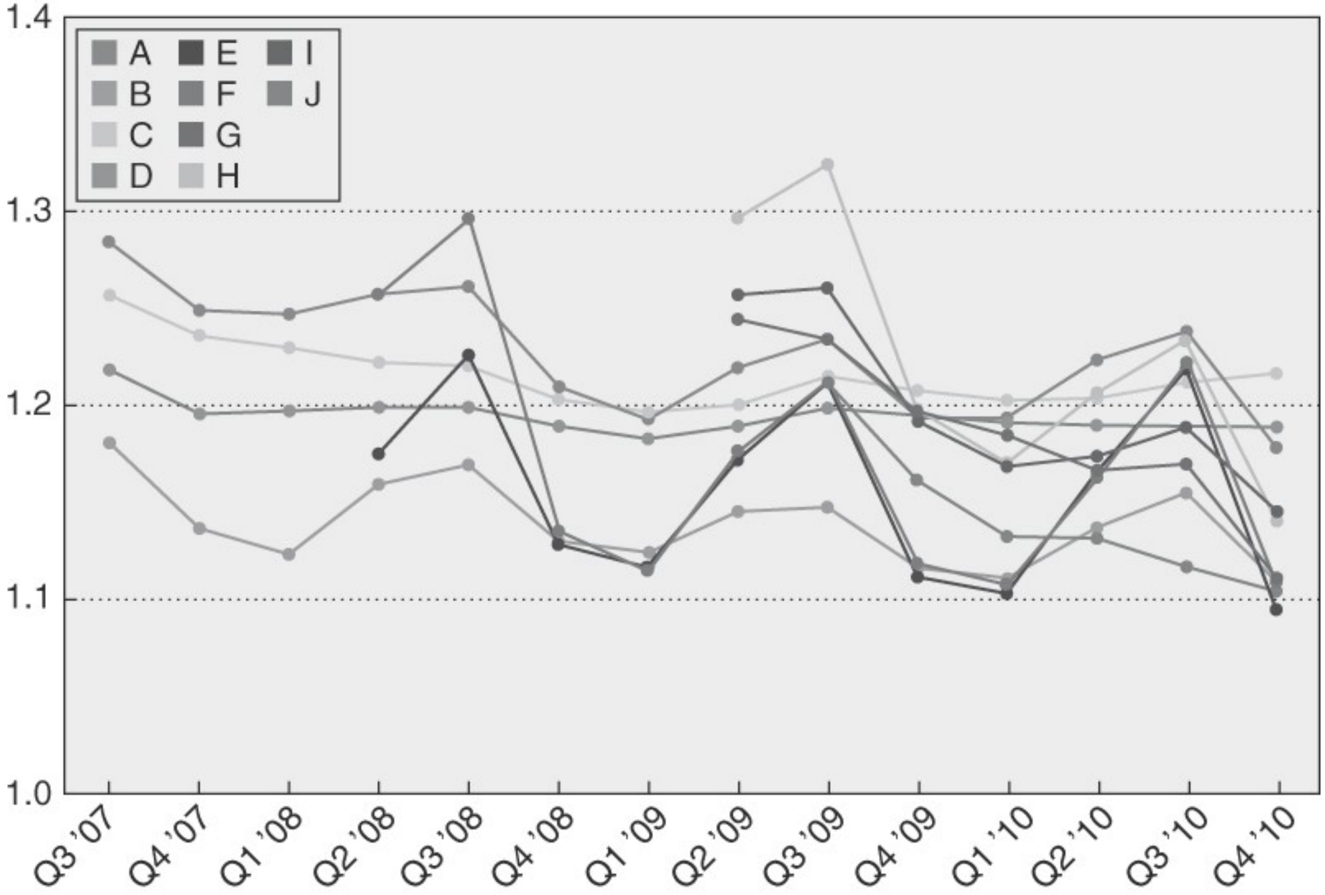
Energy consumption

- Power Utilization Effectiveness (PUE)
  - = Total facility power / IT equipment power
  - Median PUE on 2006 study was 1.69
  - Today large facilities close 1.1
- Performance
  - Latency is important metric because it is seen by users
  - Bing study: users will use search less as response time increases
  - Service Level Objectives (SLOs)/Service Level Agreements (SLAs)
    - E.g. 99% of requests be below 100 ms

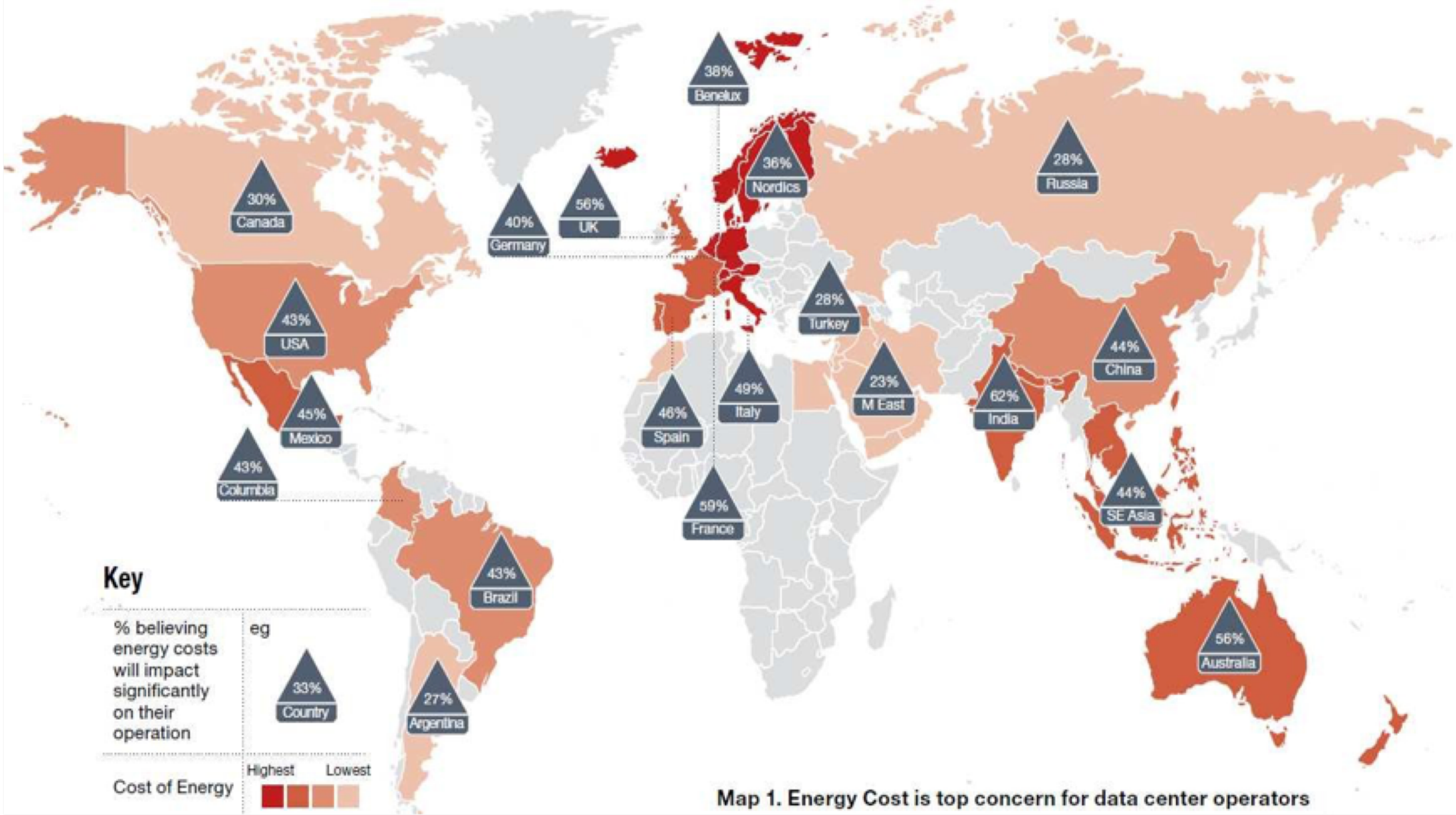
CPU Utilization and Energy Consumption



Power usage effectiveness (PUE) of 10 Google WSCs over time.



Energy Op-Ex Importance



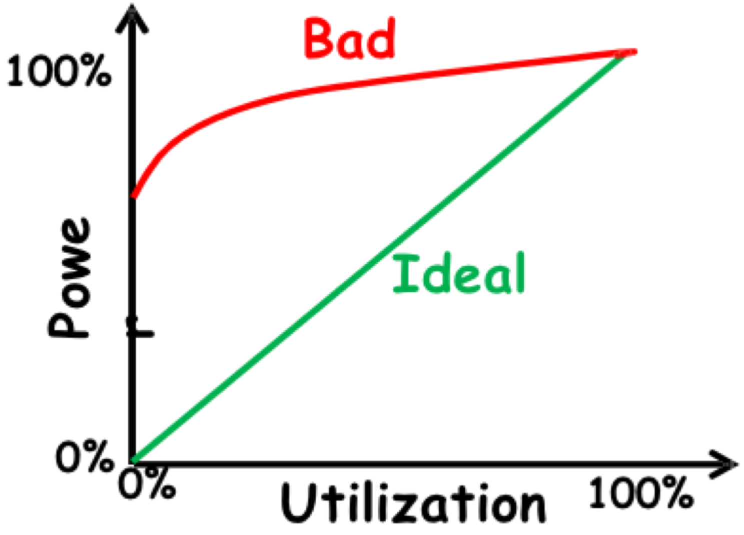
Map 1. Energy Cost is top concern for data center operators

Src: Datacenter Dynamics

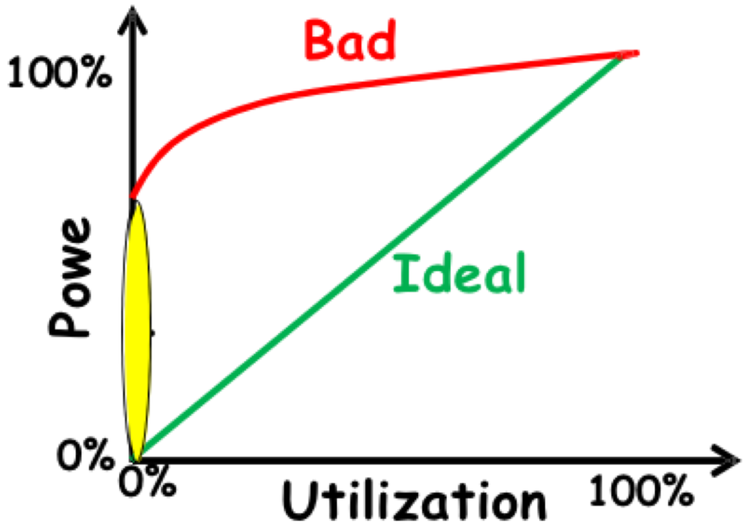
Built Energy Proportional Systems

1. No power when Idle

2. Consumption linearly increases with work (utilization) till Max Power



Cutting Idle Power

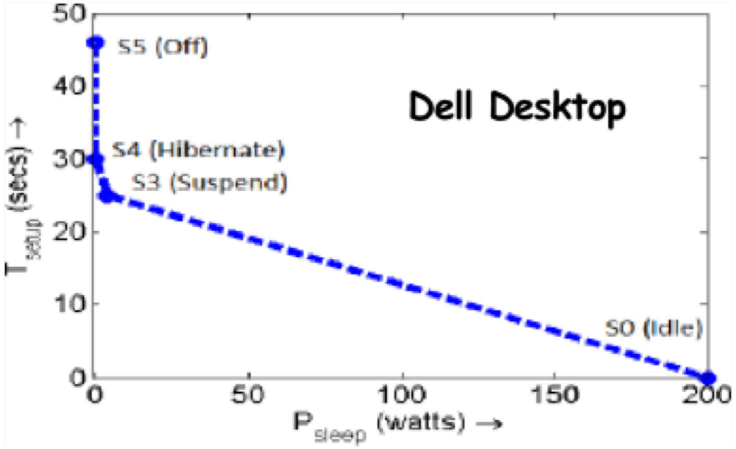


- When idle, we watour servers to be close to 0 power

- Need to ensure negligible latency to become active again

System idle low-power states

- System idle low-power states or ACPI "S-states"
  - S0 = operating state
  - S1 = CPU caches are flushed; CPUs stop executing instructions
  - S2 = CPU caches are flushed; CPUs are turned off
  - S3 = All context is lost, except for RAM; RAM & devices on
  - S4 = All context is lost (some saved to disk); RAM & devices off
  - S5 = Same as S4, except that OS doesn't save the context

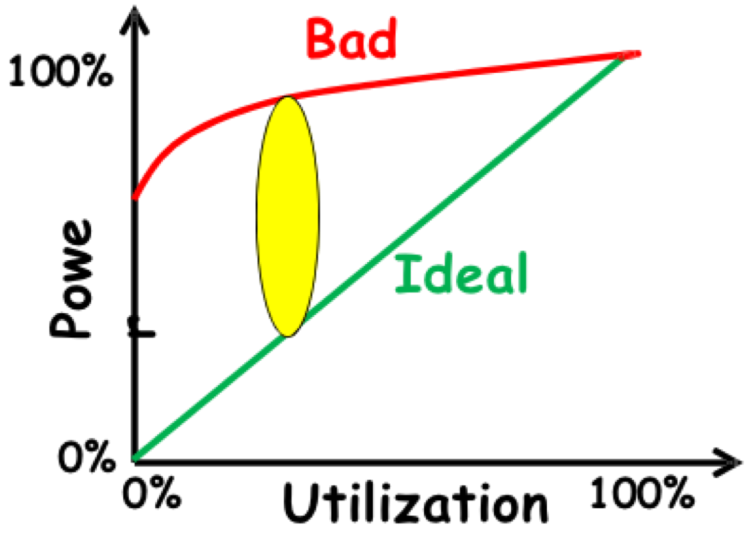


[Src: Gandhi et al., HotPower'11]

Transitions effected in software

Transitions can take significant energy and multiple secs.

Making Power Linear with Work



Need to make power consumption commensurate with the utilization (work done) by the system (each of its components)

But

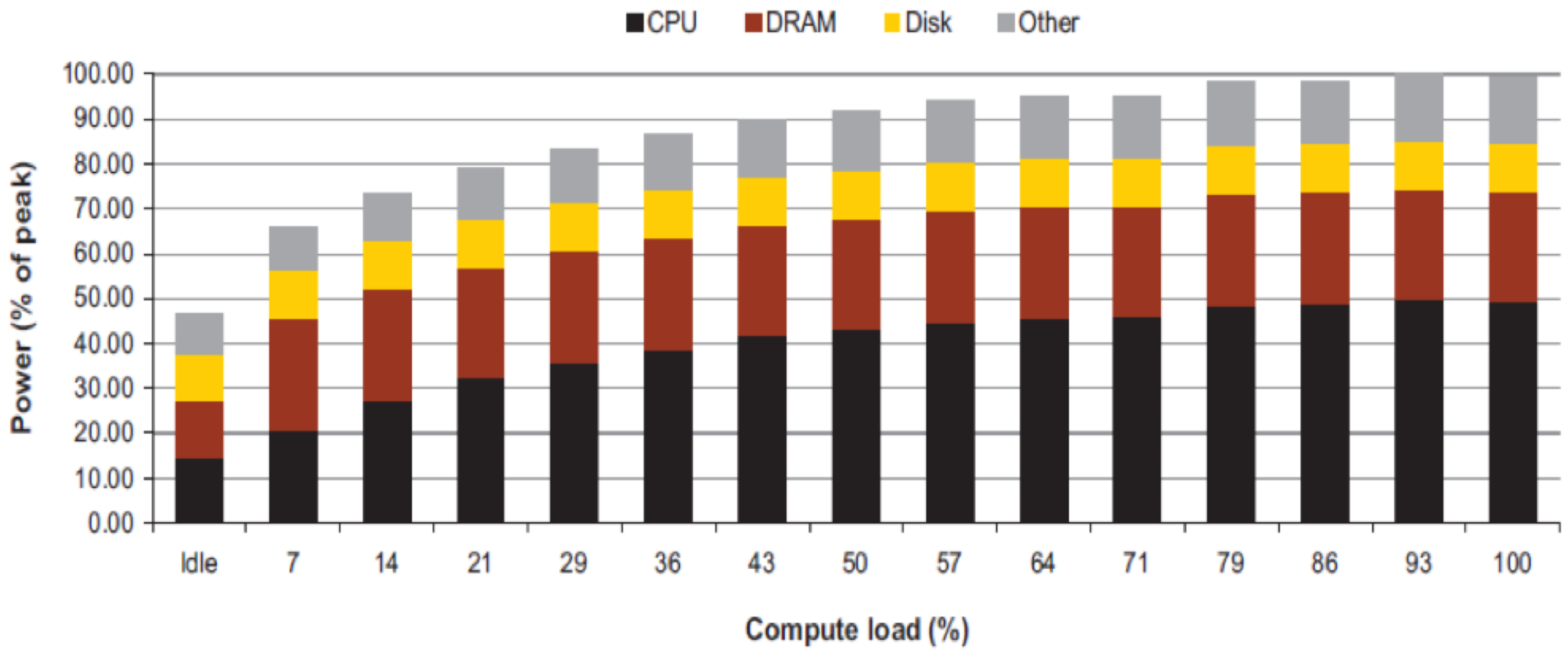
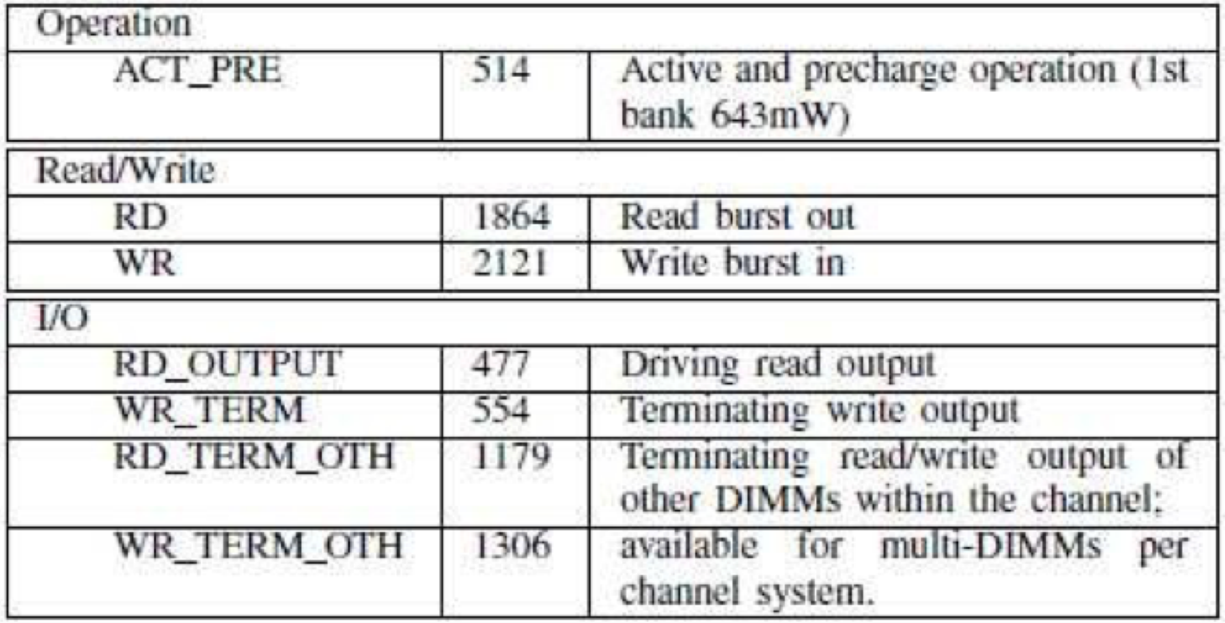


FIGURE 5.8: Subsystem power usage in an ×86 server as the compute load varies from idle to full usage.

CPU DVFS

- Reduce Power
- Reduce Energy (especially if performance is ls impacted by lower frequency)
- Power Capping
- Dynamic Thermal Management
- Continuous Monitoring and Control Loop to accommodate Dynamic Variations

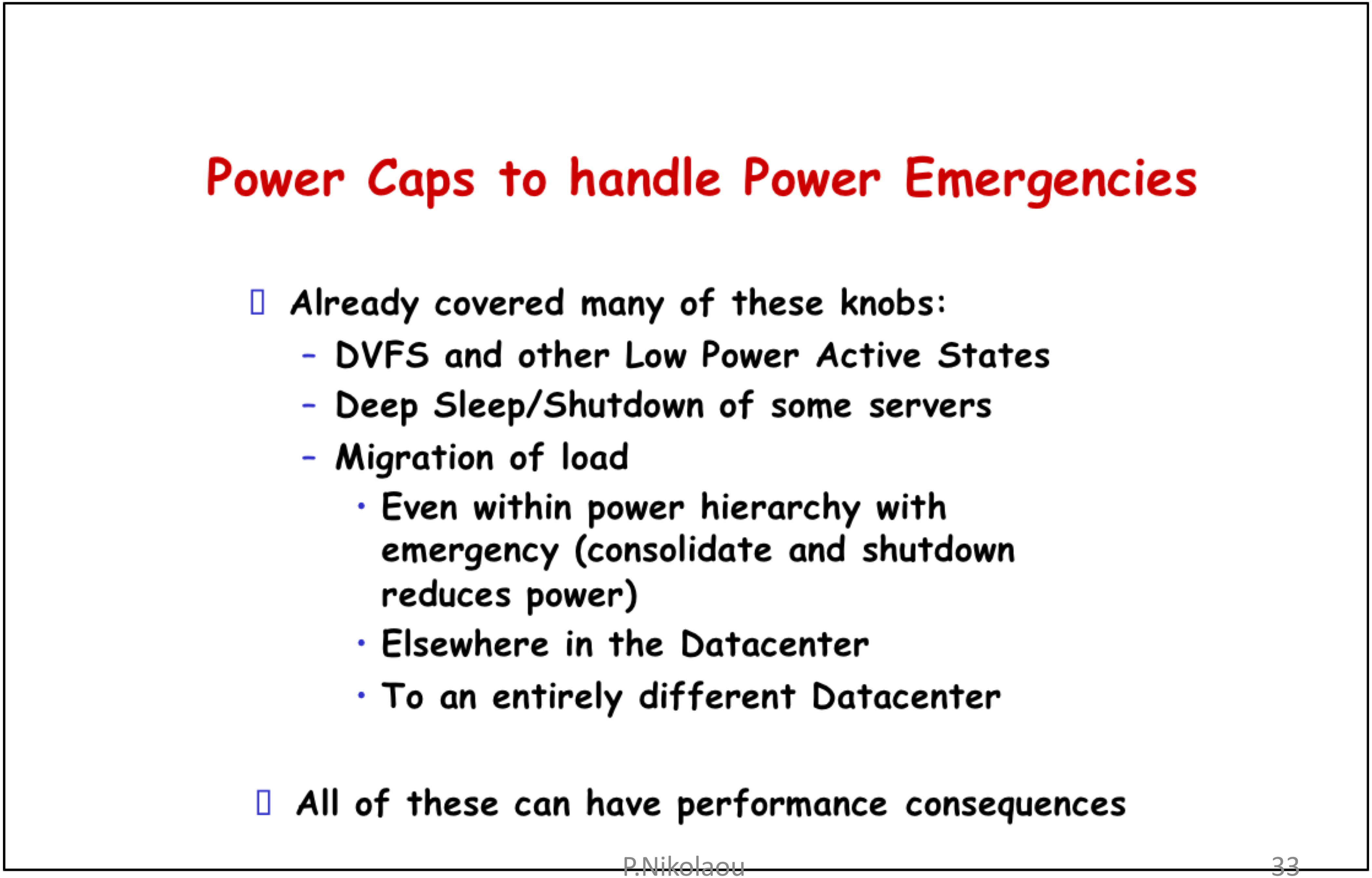
DRAM DVFS



Operation		
ACT_PRE	514	Active and precharge operation (1st bank 643mW)
Read/Write		
RD	1864	Read burst out
WR	2121	Write burst in
I/O		
RD_OUTPUT	477	Driving read output
WR_TERM	554	Terminating write output
RD_TERM_OTH	1179	Terminating read/write output of other DIMMs within the channel;
WR_TERM_OTH	1306	available for multi-DIMMs per channel system.

Dynamic On-die Termination to adjust impedence for signal integrity

- Memory active low-power modes [Deng et al., ASPLOS'11]
  - DVFS of the memory controller and PLL
  - DFS of the memory bus, DIMM interface, and DRAM devices

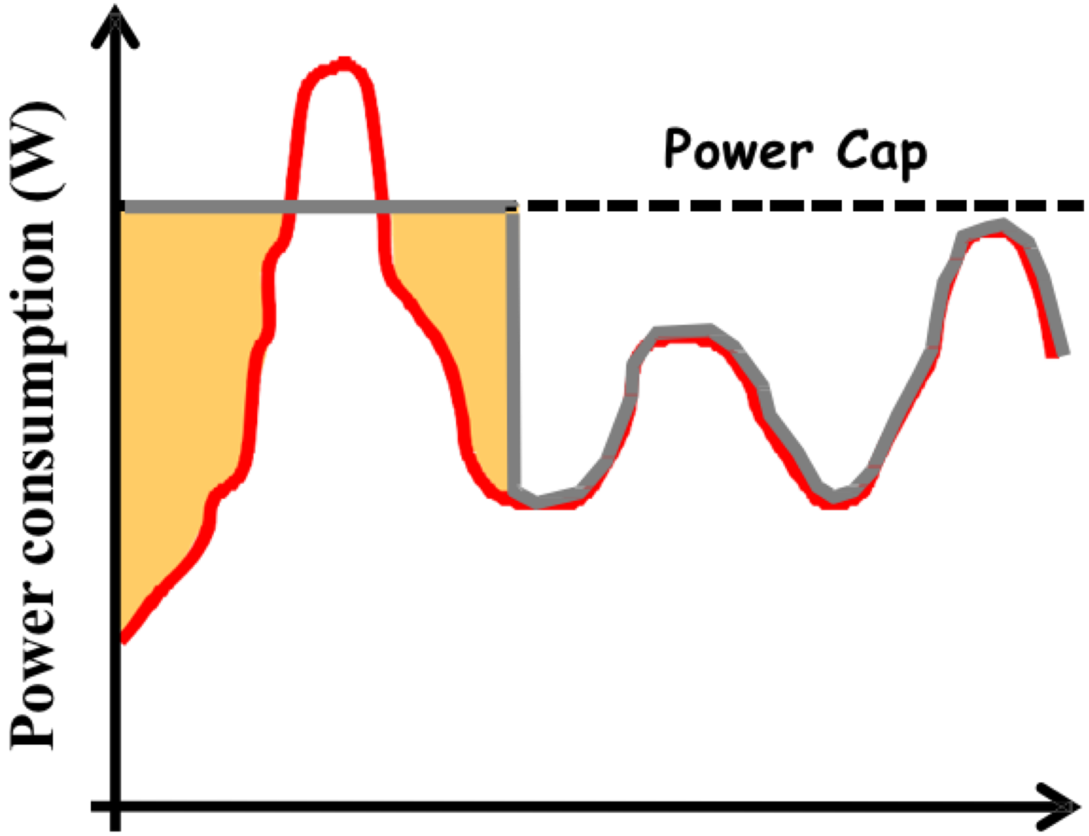


Power Caps to handle Power Emergencies

- Already covered many of these knobs:
  - DVFS and other Low Power Active States
  - Deep Sleep/Shutdown of some servers
  - Migration of load
    - Even within power hierarchy with emergency (consolidate and shutdown reduces power)
    - Elsewhere in the Datacenter
    - To an entirely different Datacenter

- All of these can have performance consequences

Power Capping Approach



Inside Google's Datacenter

Watch Video:

https://www.youtube.com/watch?v=XZmGGAbHqa0

Total Cost of Ownership (TCO)

- Key optimization metric
- Capture both capex and opex
- Capital Expenses: land, building, servers, switches, power, cooling, sw licences
- Operational Expenses: energy, spares, maintenance, personnel

Why TCO is important?

Data center research became very important

Large industry with economic and society impact

Environmental impact

Big investment

Main target of all data center research is to finally reduce the TCO and increase profit

Energy reduction

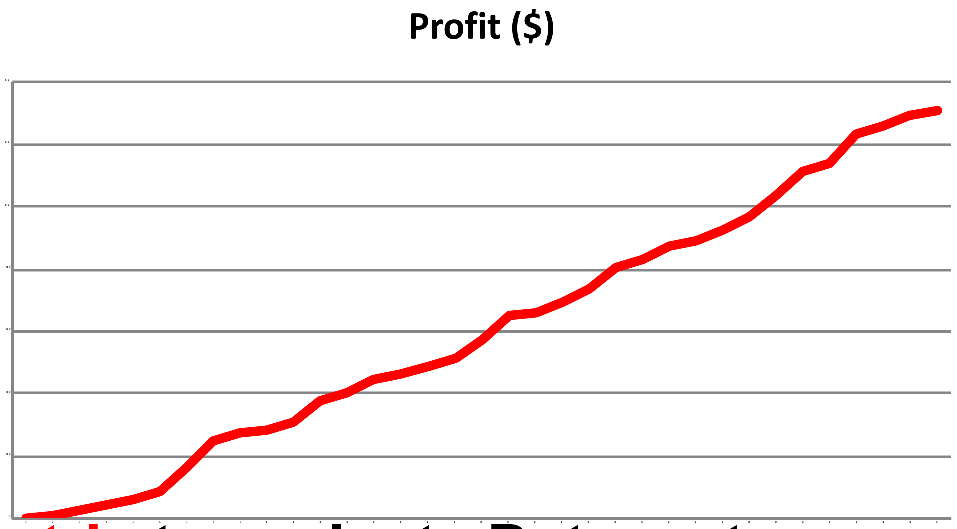
Power efficiency

Green data centers

Efficient maintenance

New server designs

etc.



This makes TCO the primary metric to evaluate Datacenters.

TCO tools

Tools that main purpose is to calculate TCO

Can be used for assessing Datacenter design trade-offs through design space exploration.

TCO tools can be useful both for

1) research - estimate how new server designs, power management techniques and etc. affect TCO

2) enterprise environment – plan Datacenter, monitor monthly Datacenter costs

TCO tools are important because they allow you to assess the most important Datacenter efficiency metric which is off-course TCO.

Overview of existing TCO tools

Public available TCO tools

- Spreadsheet based tools
  - James Hamilton spreadsheet
  - True TCO calculator by Uptime Institute
- Web based tools
  - Calculators for calculating TCO savings on specific company’s products
  - Usually companies use them to communicate the benefits of their solution
- Research Tools
  - Developed to evaluate different case studies

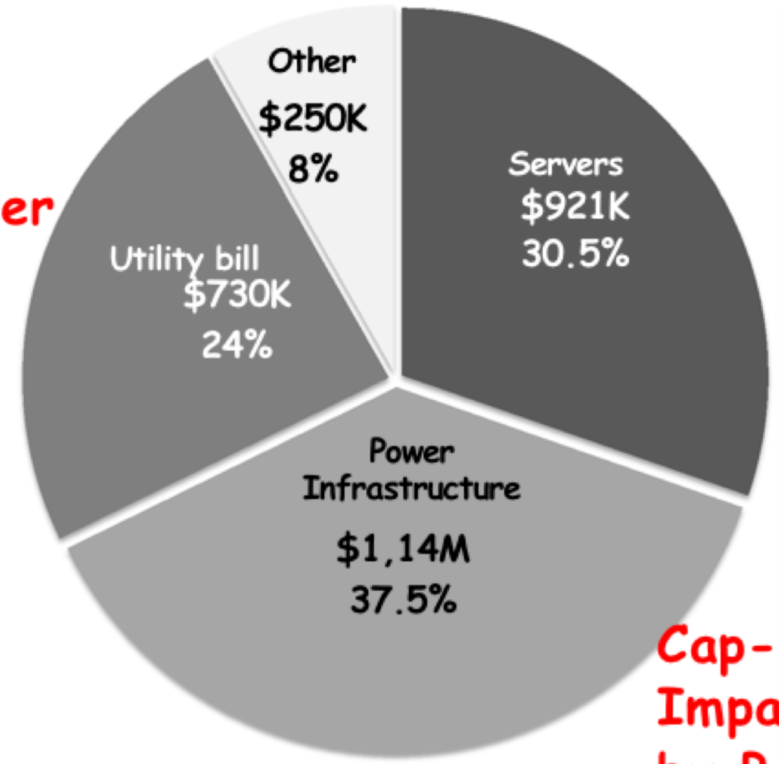
In-house models

- Academic tools
- Facebook, Google, IBM and for sure many other companies

Companies that provide TCO services

TCO

Op-Ex
Impacted by
Energy & Power



Cap-Ex
Impacted mainly
by Power

Assumption: 20,000 servers, 1.5 PUE, 15$/W Cap-ex, *4yr server & 12 yr infrastructure amortization (Tier-2)*

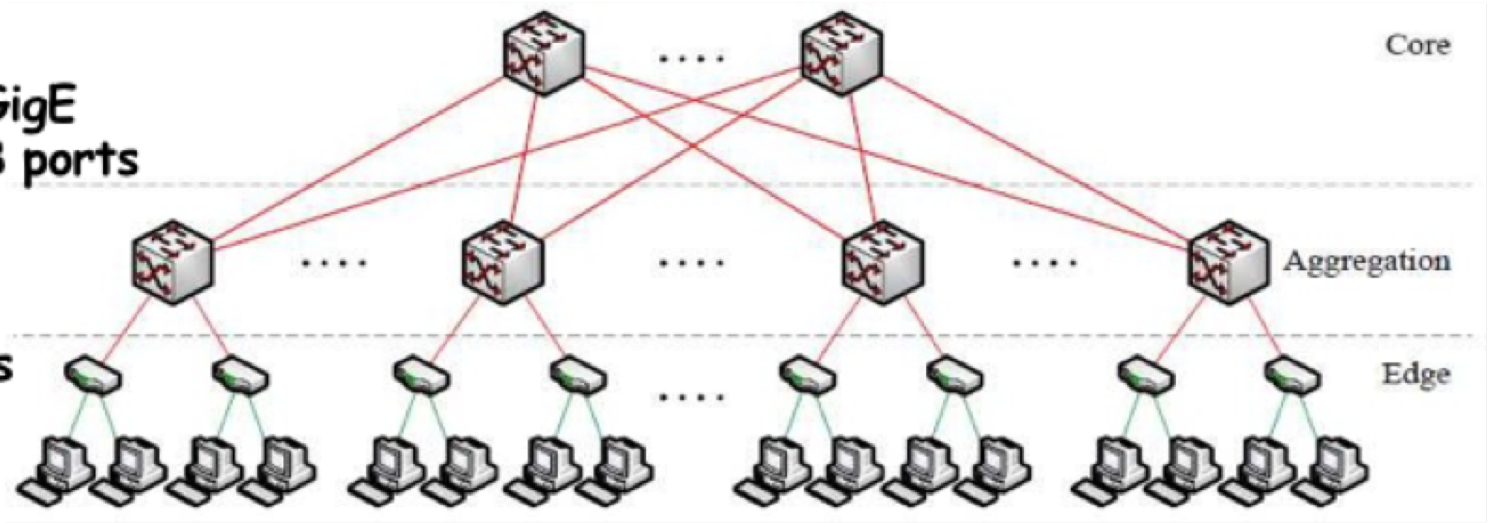
Datacenter Networks

Conventional Hierarchical Design

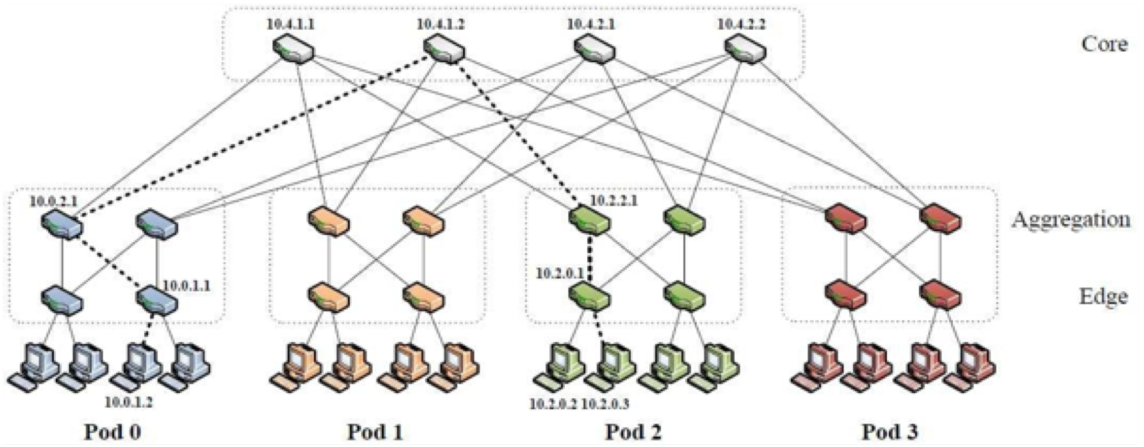
[Src:Al Fares et al., Sigcomm'08]

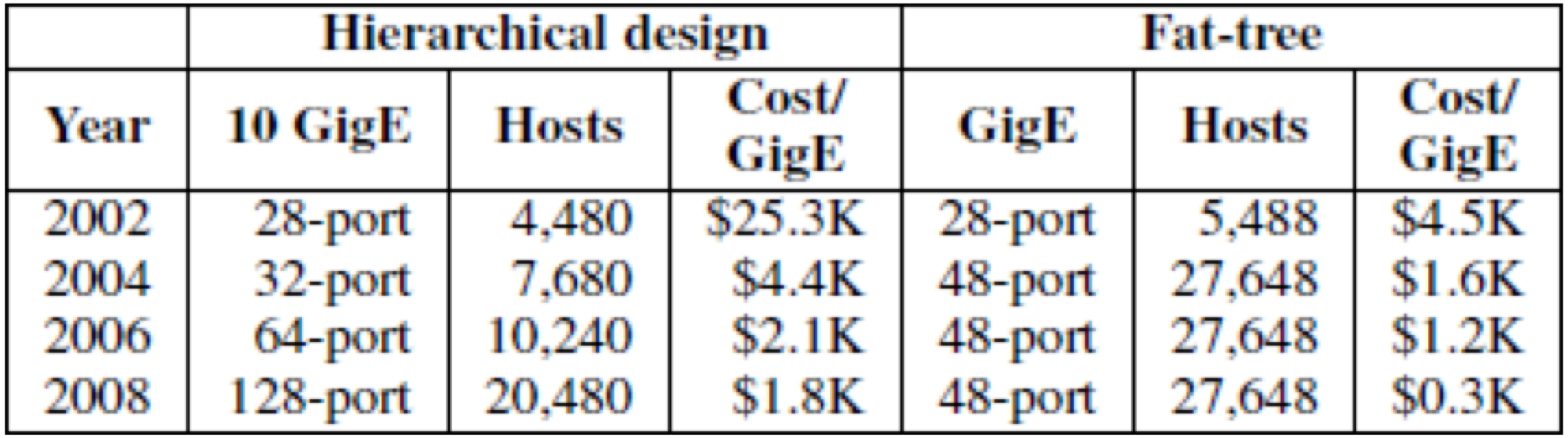
10GigE
128 ports

GigE
48 ports



Fat-Tree Design with Wimpy Switches





	Hierarchical design			Fat-tree		
Year	10 GigE	Hosts	Cost/ GigE	GigE	Hosts	Cost/ GigE
2002	28-port	4,480	$25.3K	28-port	5,488	$4.5K
2004	32-port	7,680	$4.4K	48-port	27,648	$1.6K
2006	64-port	10,240	$2.1K	48-port	27,648	$1.2K
2008	128-port	20,480	$1.8K	48-port	27,648	$0.3K

[Src: Al Fares et al., Sigcomm'08]

Services Provided by Datacenters

Google
amazon.com
NETFLIX
Microsoft® Office 365
Spotify
Google docs
bing
facebook
salesforce

These things are really big

Google	100 billion searches per month
facebook	1.15 billion users
amazon.com	120+ million users

Services Provided by Datacenters

- **Online Workload:** Web Search
- High level view
  - Clients
  - Front-end
  - Index
  - Document Servers
- Index can be huge
- Highly partitioned and replicated
- Metric of interest (Quality of Service QoS)
  - Average response time but also tail latency
  - 99th % less than 100s ms

Offline Workload

- Batch processing framework: MapReduce
  - **Map:** applies a programmer-supplied function to each logical input record
    - Runs on thousands of computers
    - Provides new set of key-value pairs as intermediate values
  - **Reduce:** collapses values using another programmer-supplied function

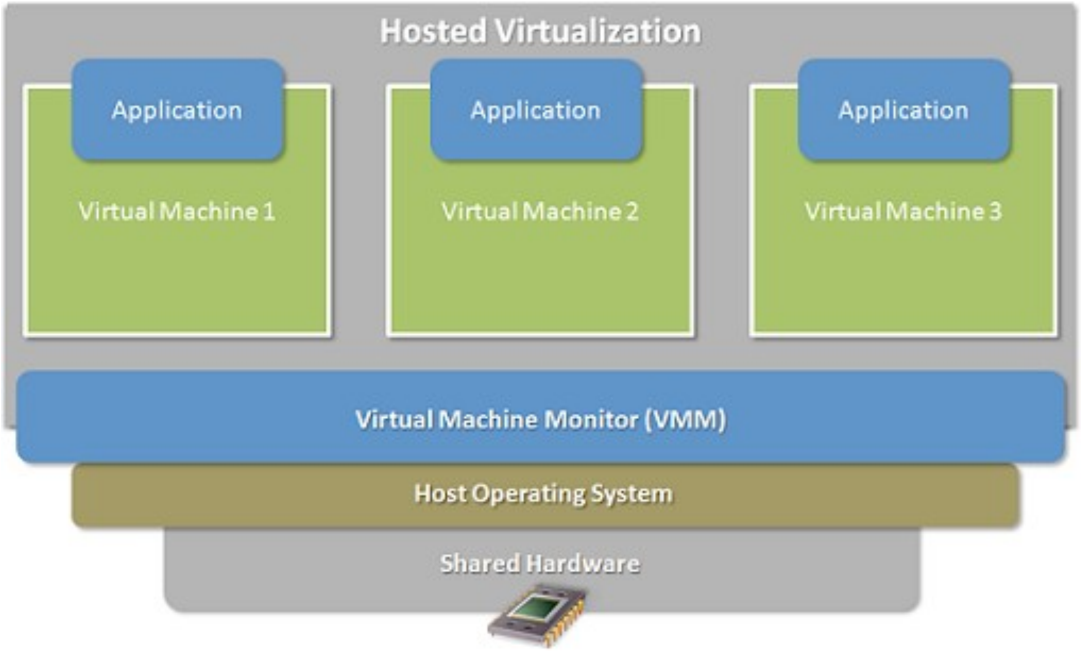
Map Reduce

- Example:
  - map (String key, String value):
    - // key: document name
    - // value: document contents
    - for each word w in value
      - EmitIntermediate(w,"1"); // Produce list of all words
  - reduce (String key, Iterator values):
    - // key: a word
    - // value: a list of counts
    - int result = 0;
    - for each v in values:
      - result += ParseInt(v); // get integer from key-value pair
    - Emit(AsString(result));

Cloud Computing

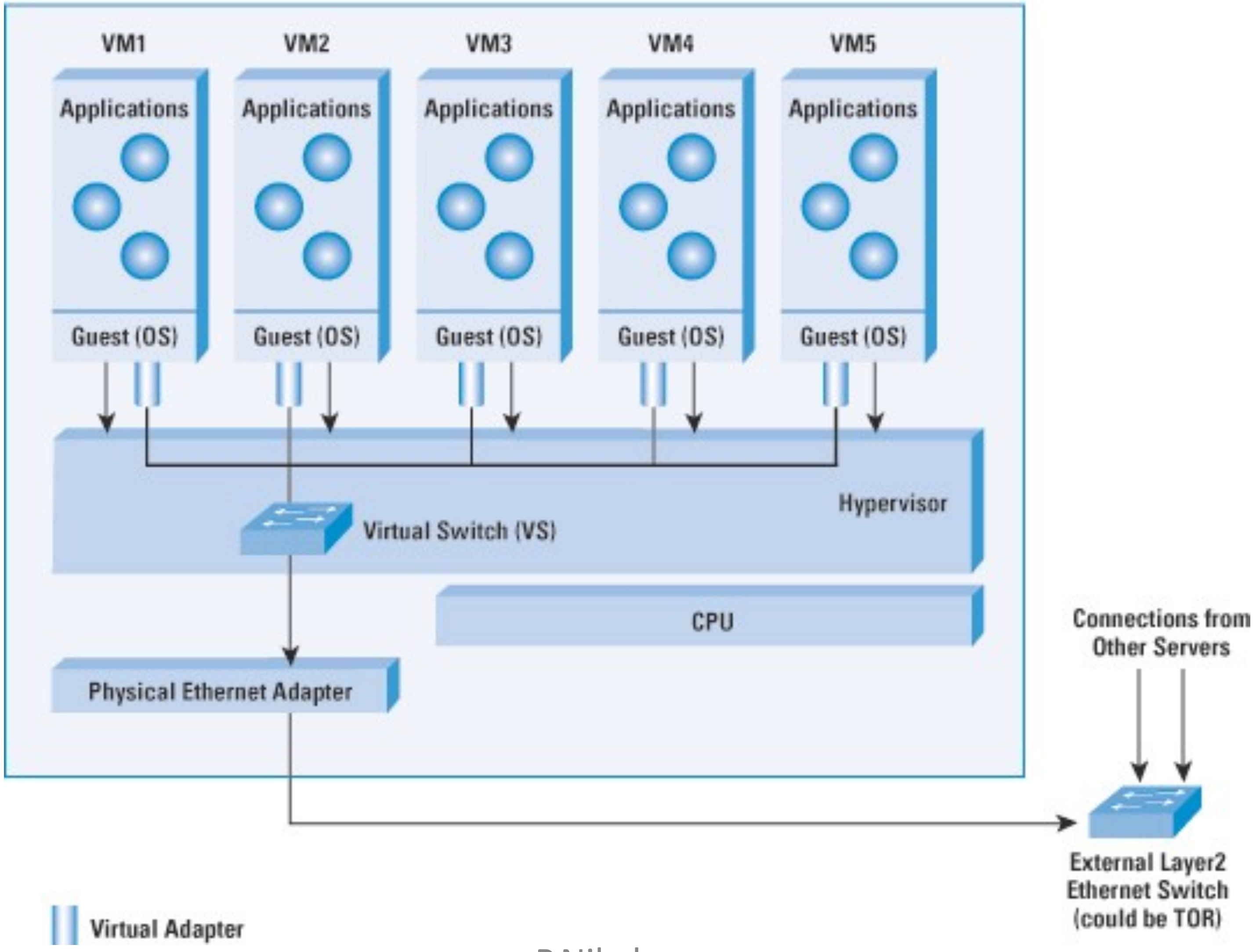
- WSCs offer economies of scale that cannot be achieved with a datacenter:
  - 5.7 times reduction in storage costs
  - 7.1 times reduction in administrative costs
  - 7.3 times reduction in networking costs
  - This has given rise to cloud services such as Amazon Web Services
    - “Utility Computing”
    - Based on using open source virtual machine and operating system software

Host Virtualization

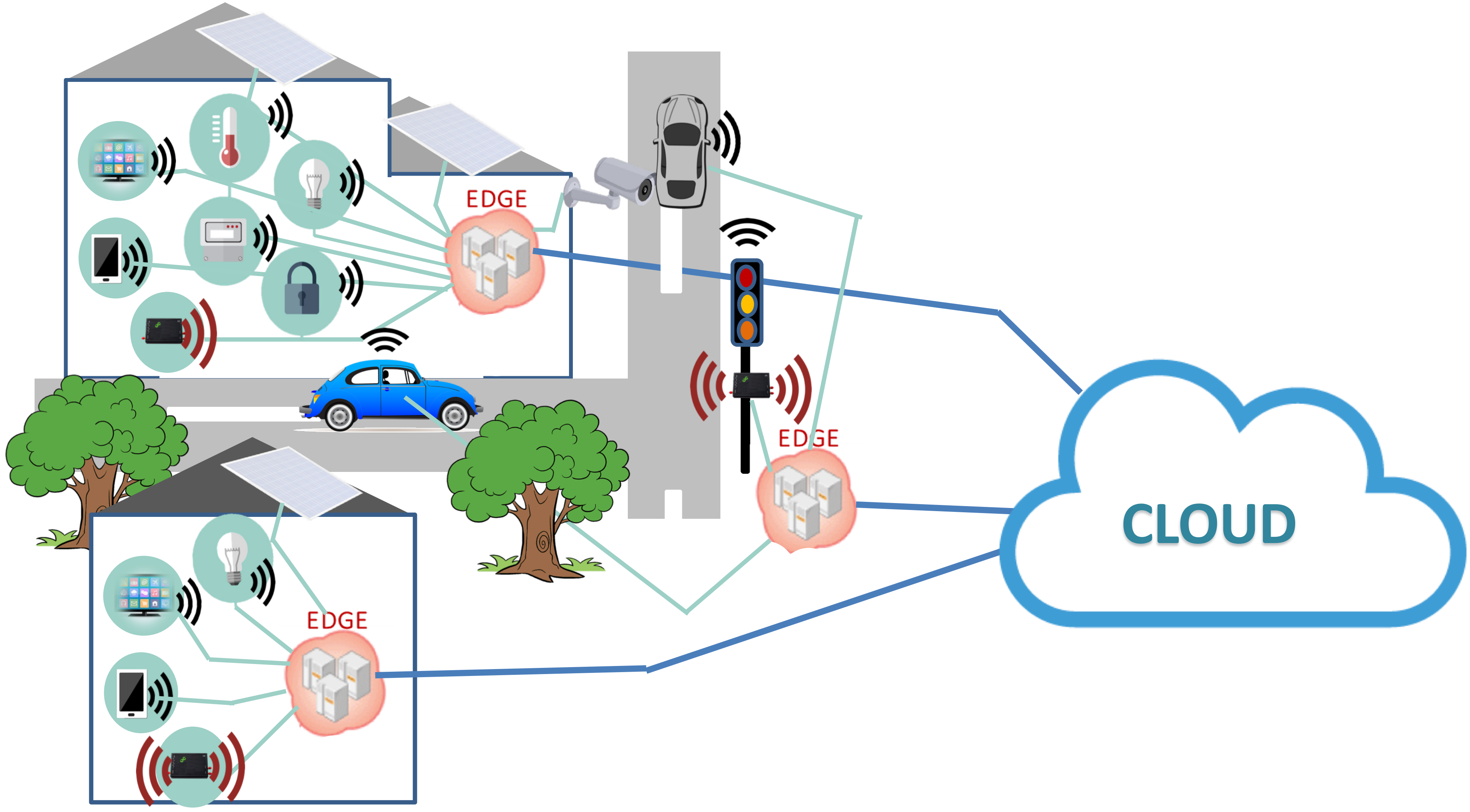


- Multiple virtual machines on one physical machine
- Applications run unmodified as on real machine
- VM can migrate from one computer to another

VMM Virtual Switches

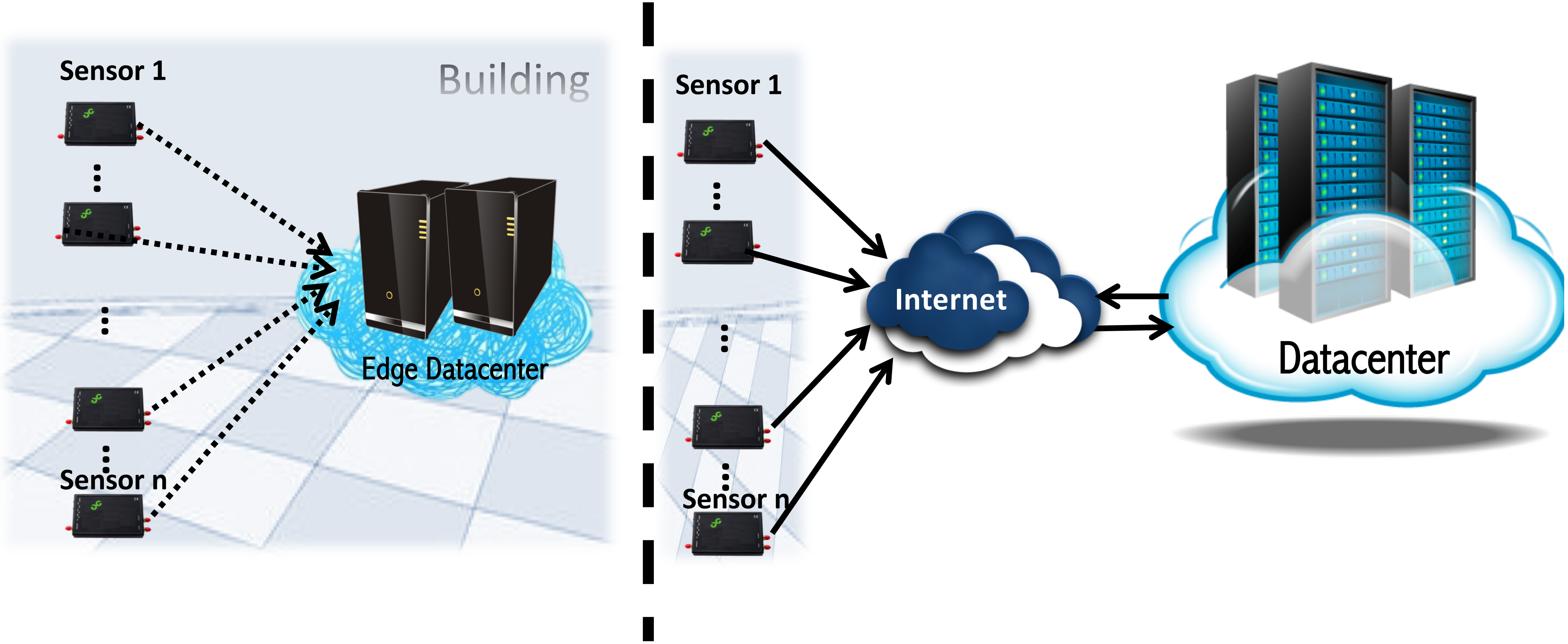


Edge Vs Cloud Datacenters



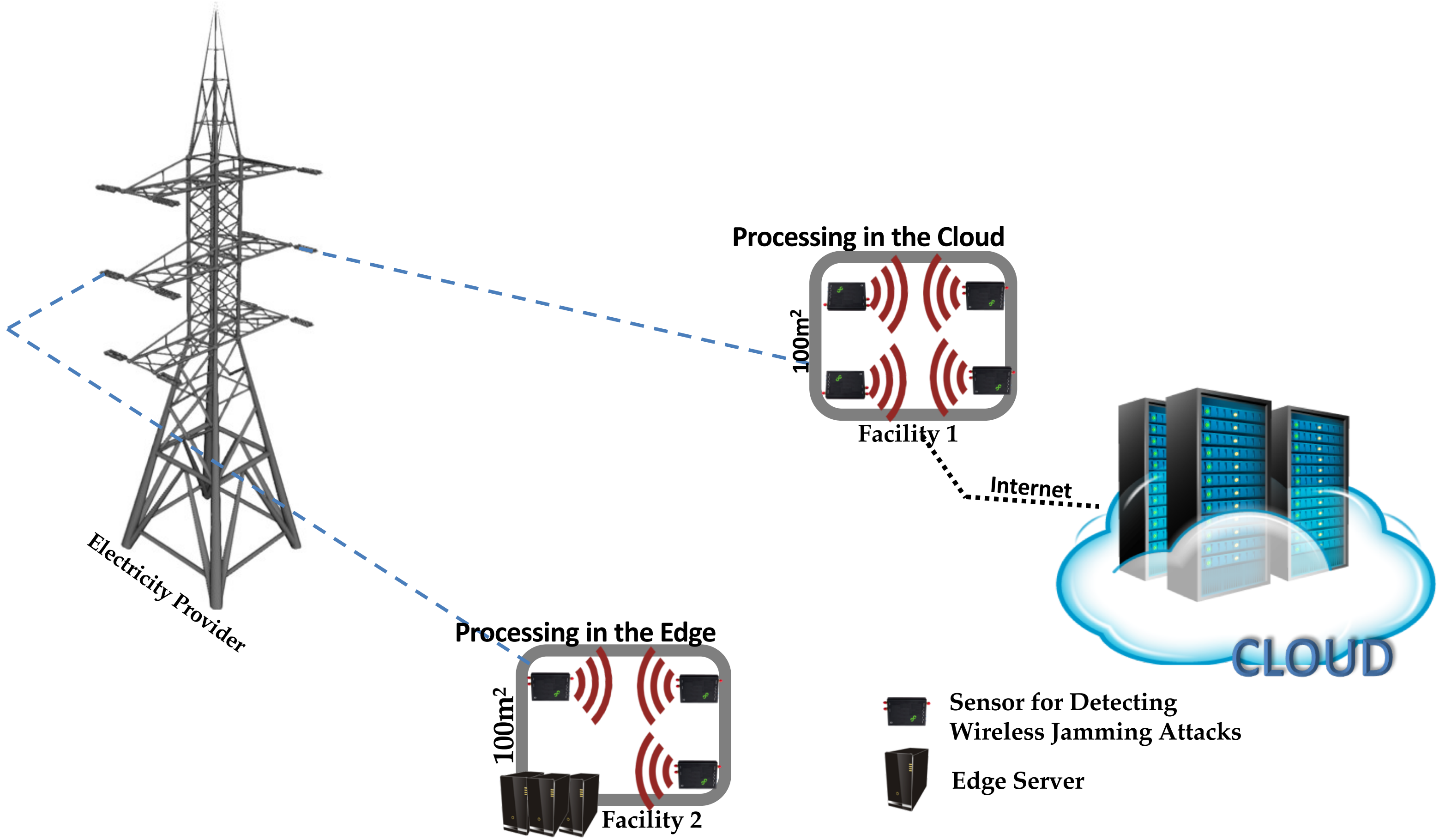
P.Nikolaou

Edge Vs Cloud Architecture

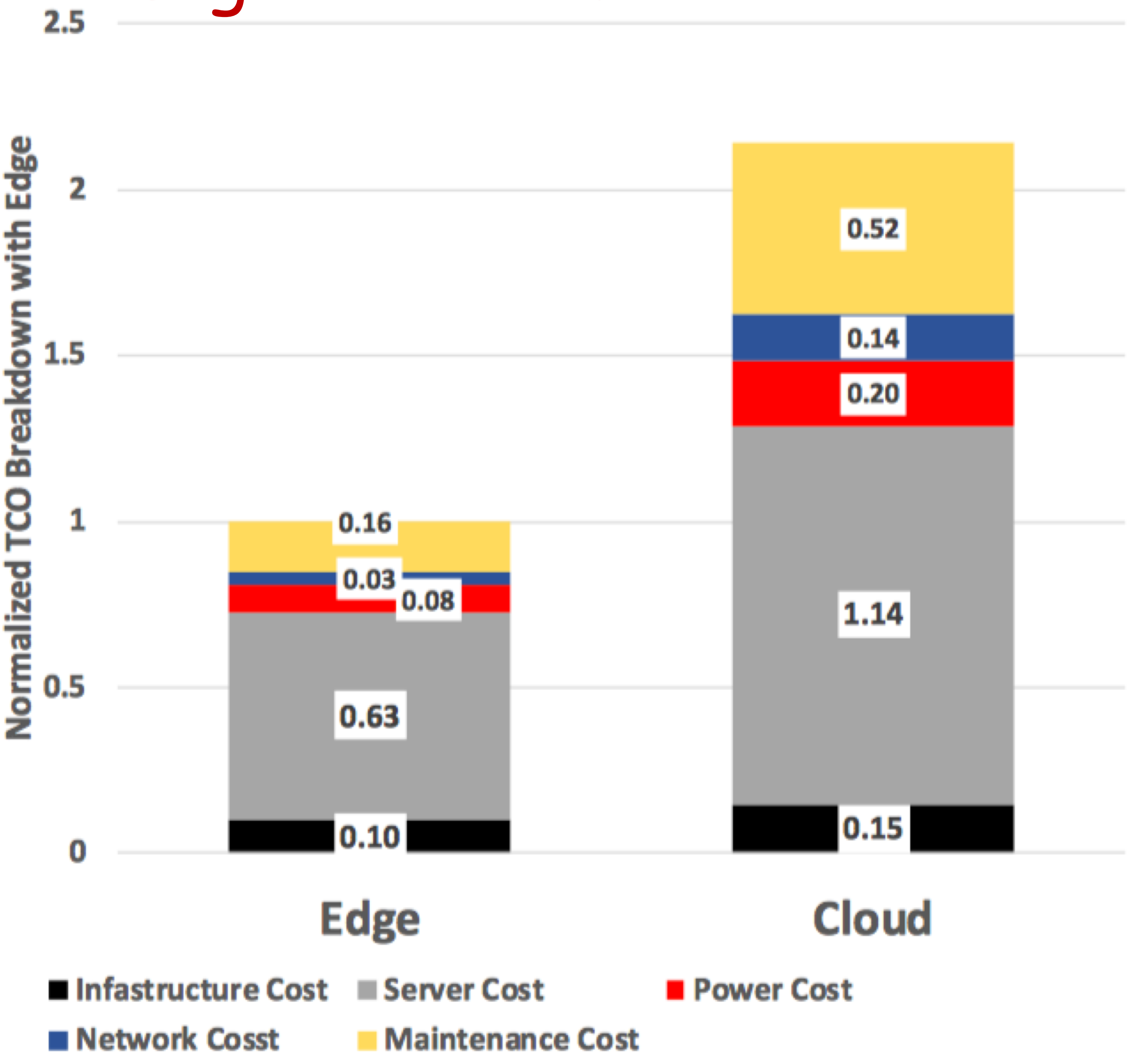


P.Nikolaou

Edge Vs Cloud Datacenters

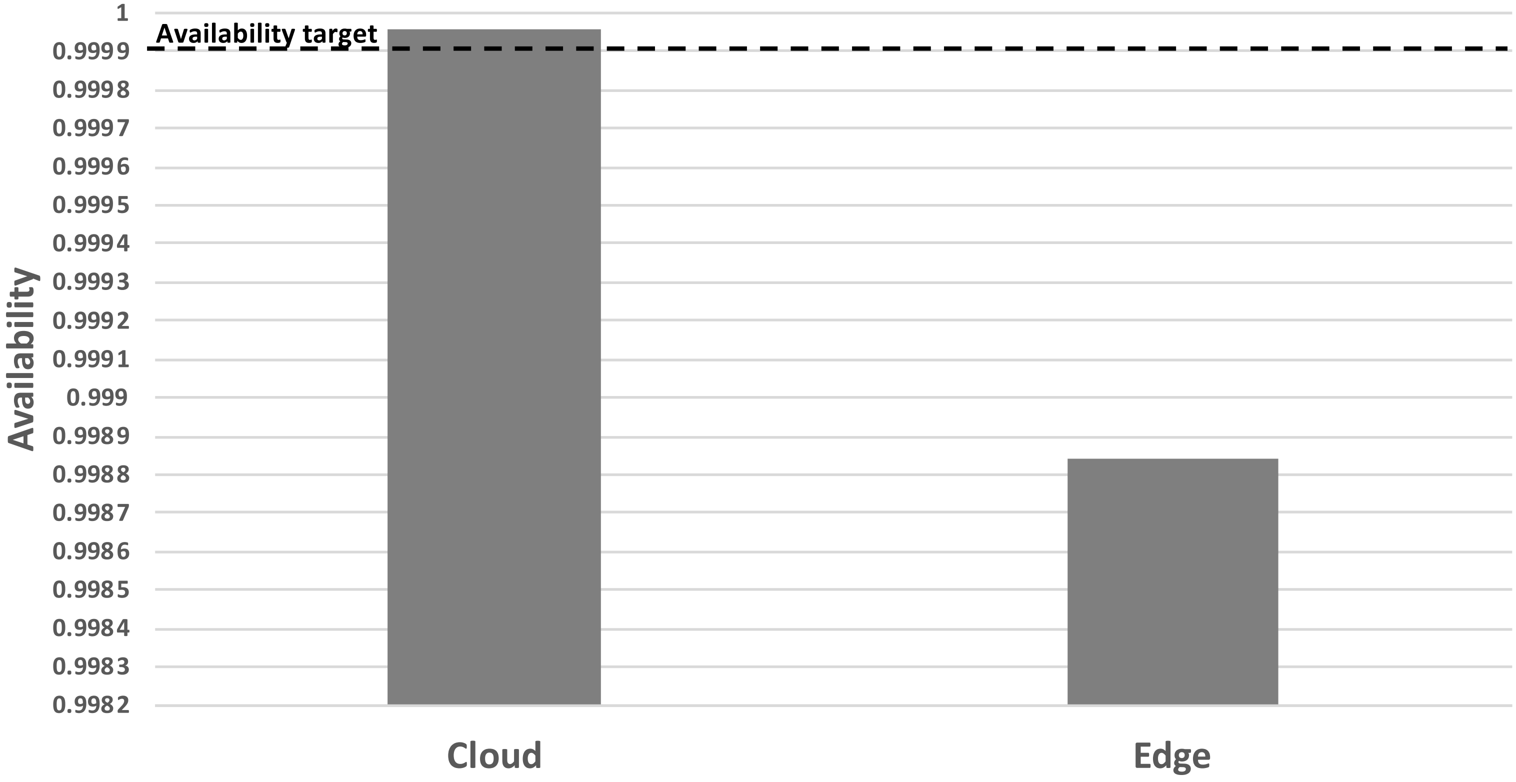


Edge Vs Cloud Datacenters

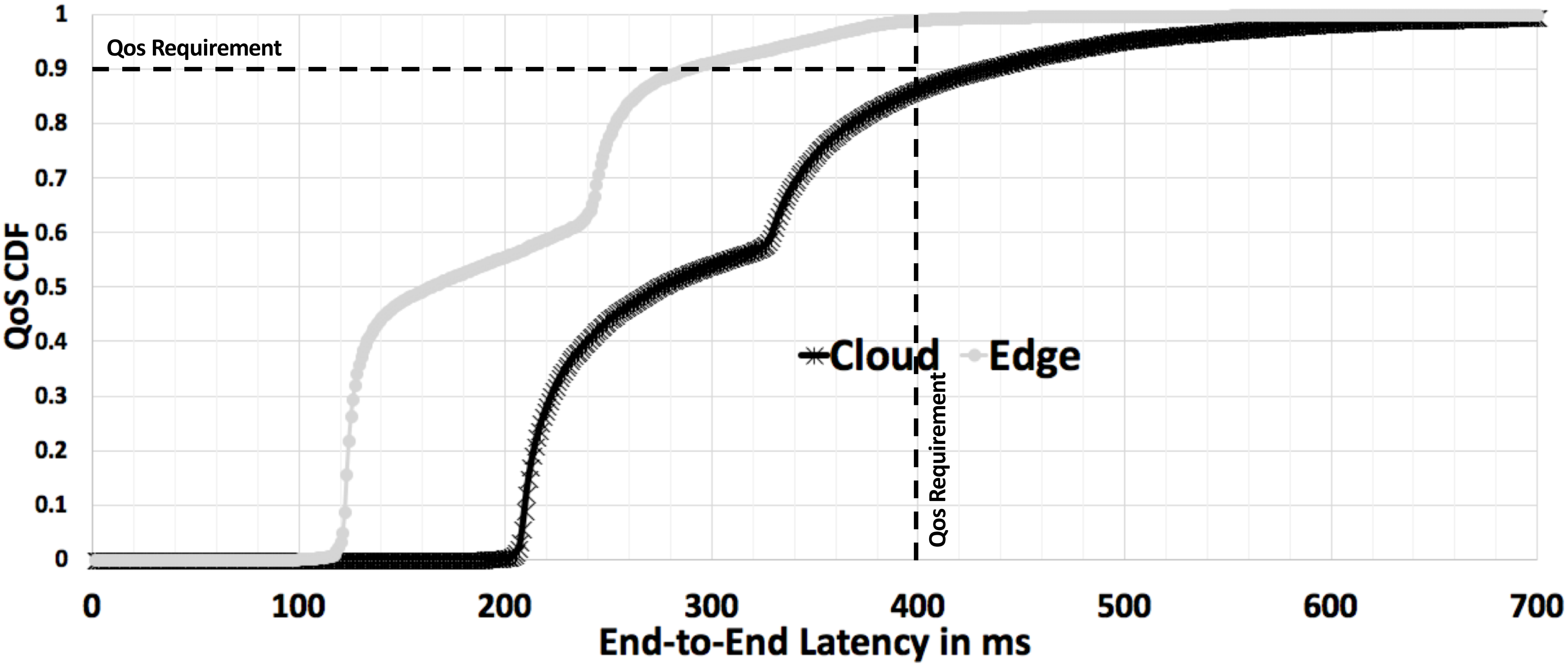


P.Nikolaou

Edge Vs Cloud Availability

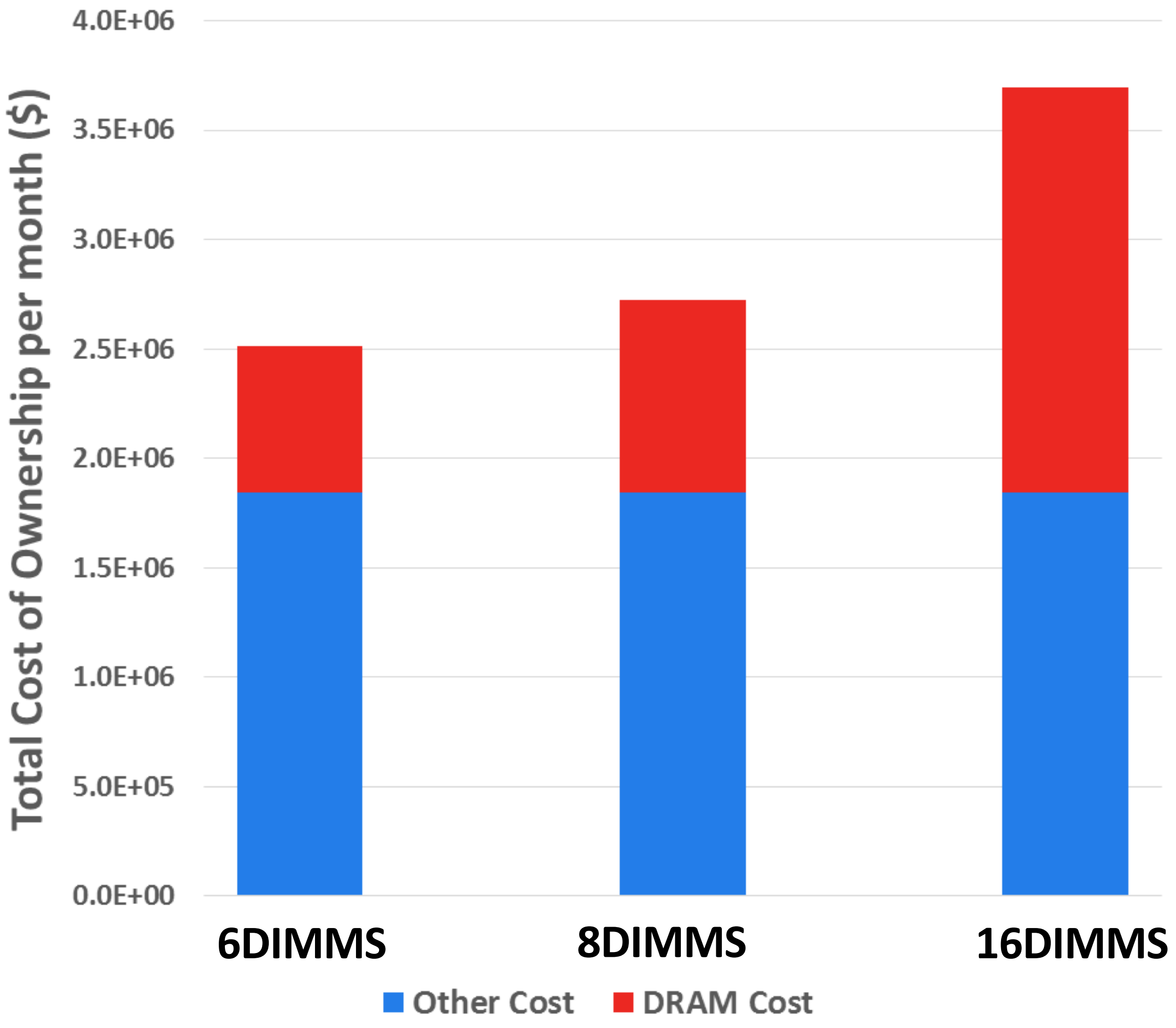


Edge Vs Cloud End to End Latency



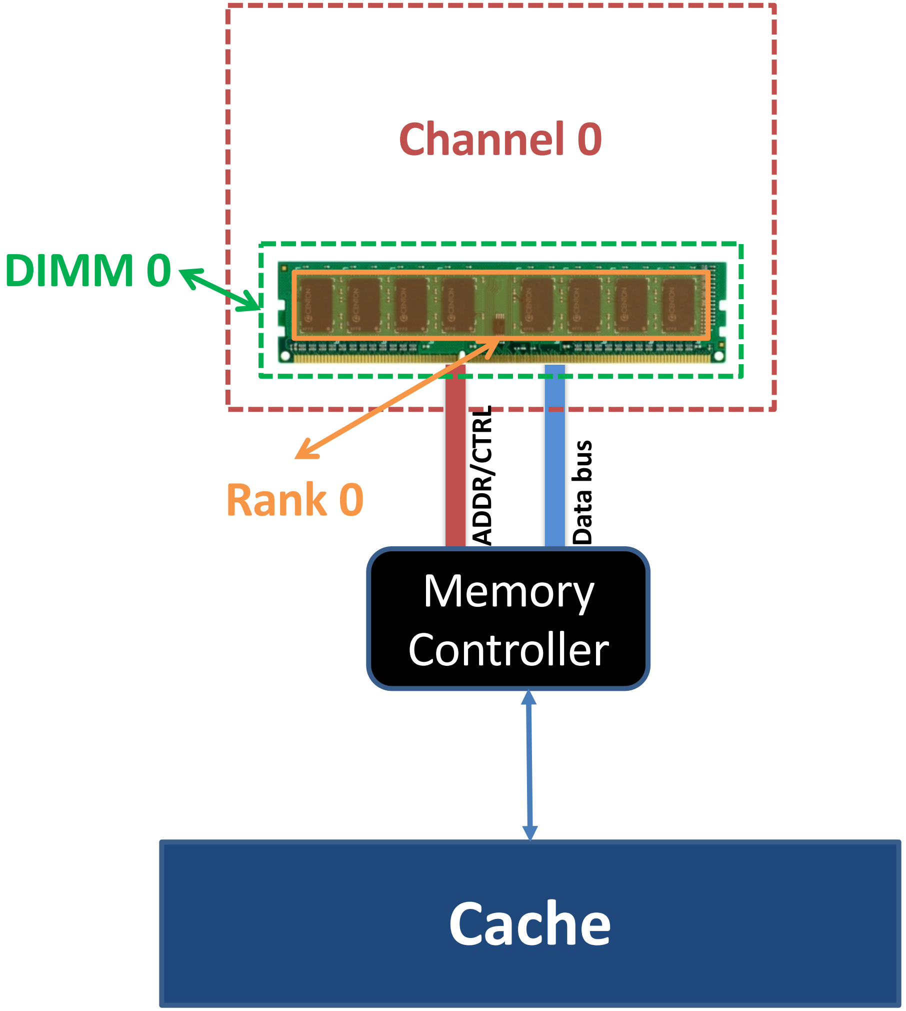
DRAM Architecture

DRAM Contribution on the TCO



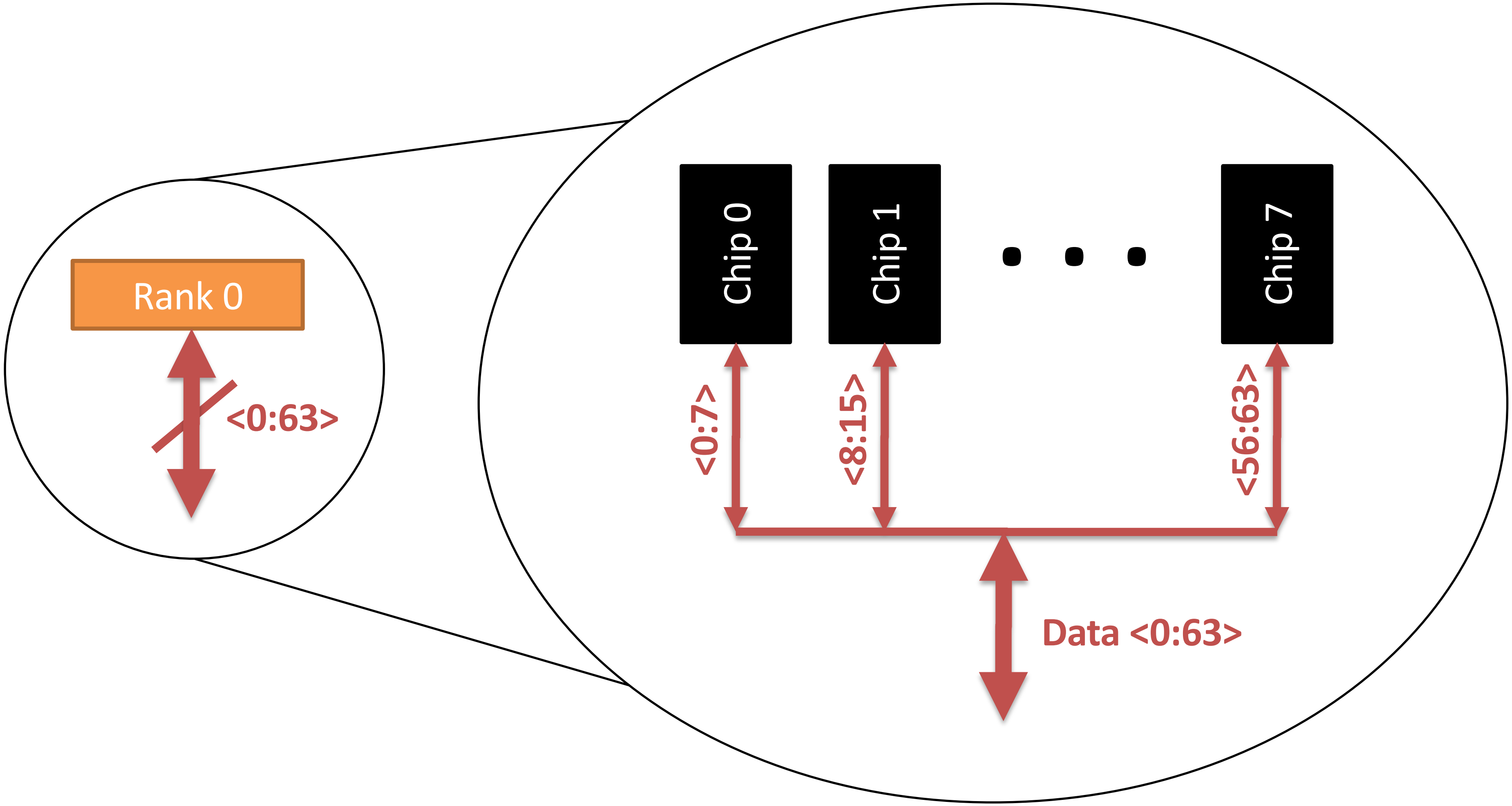
DRAM Cost is Significant!!

DRAM array Organization



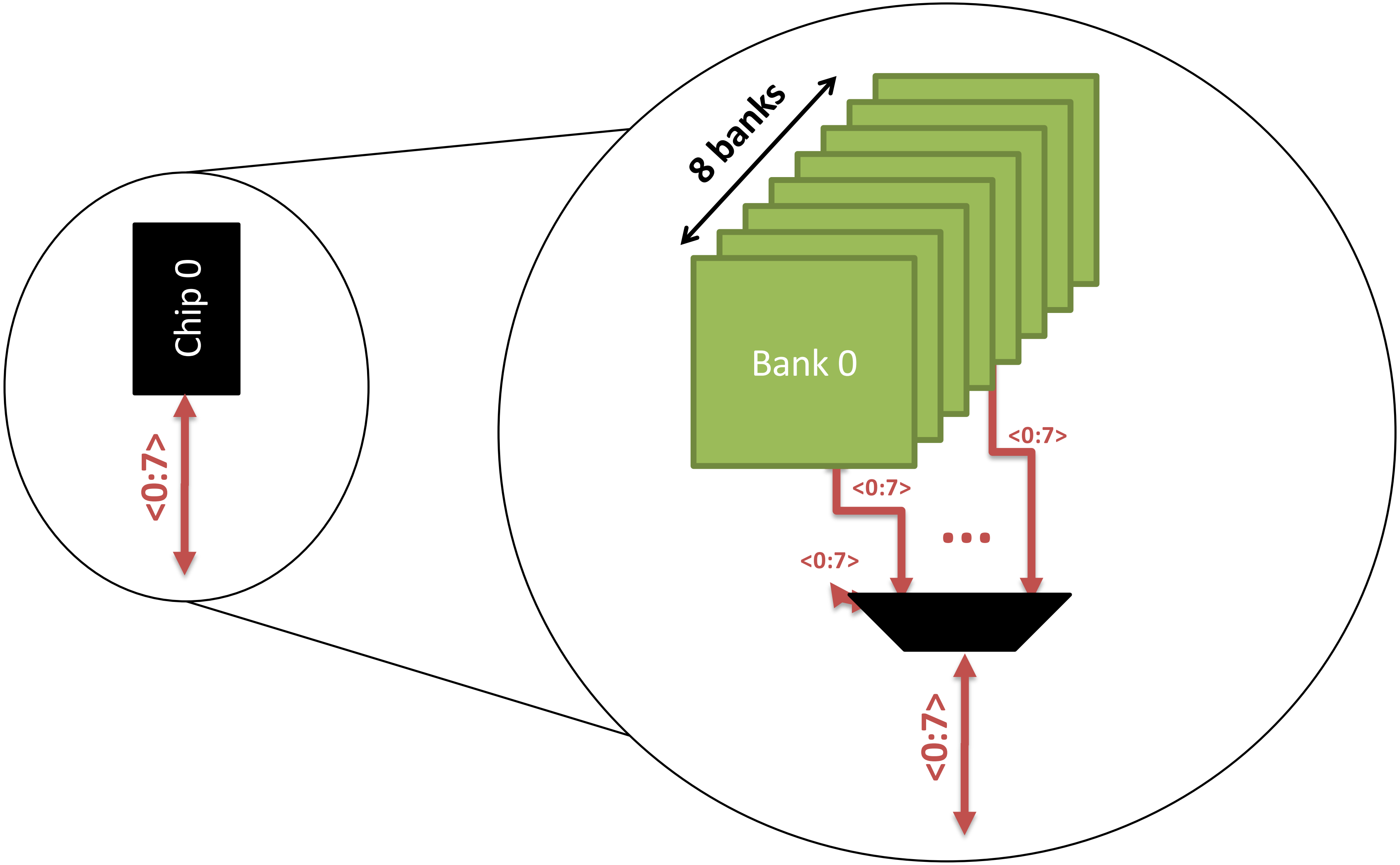
Based on "Scalable Many-Core Memory Systems ", Onur Mutlu, ACACES 2013

DRAM array Organization



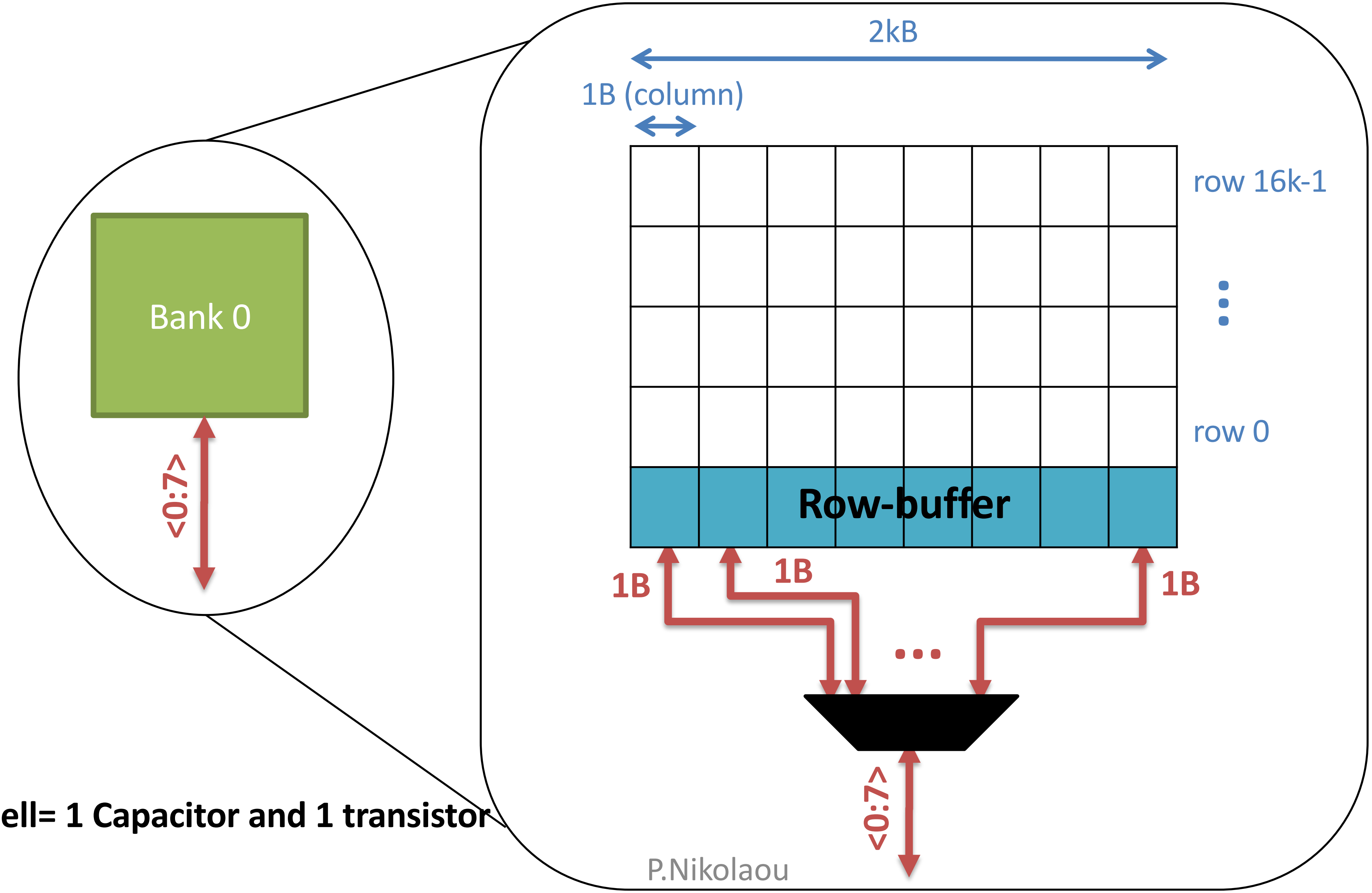
Based on “Scalable Many-Core Memory Systems ”, Onur Mutlu, ACACES 2013

DRAM array Organization



Based on "Scalable Many-Core Memory Systems", Onur Mutlu, ACACES 2013

DRAM array Organization



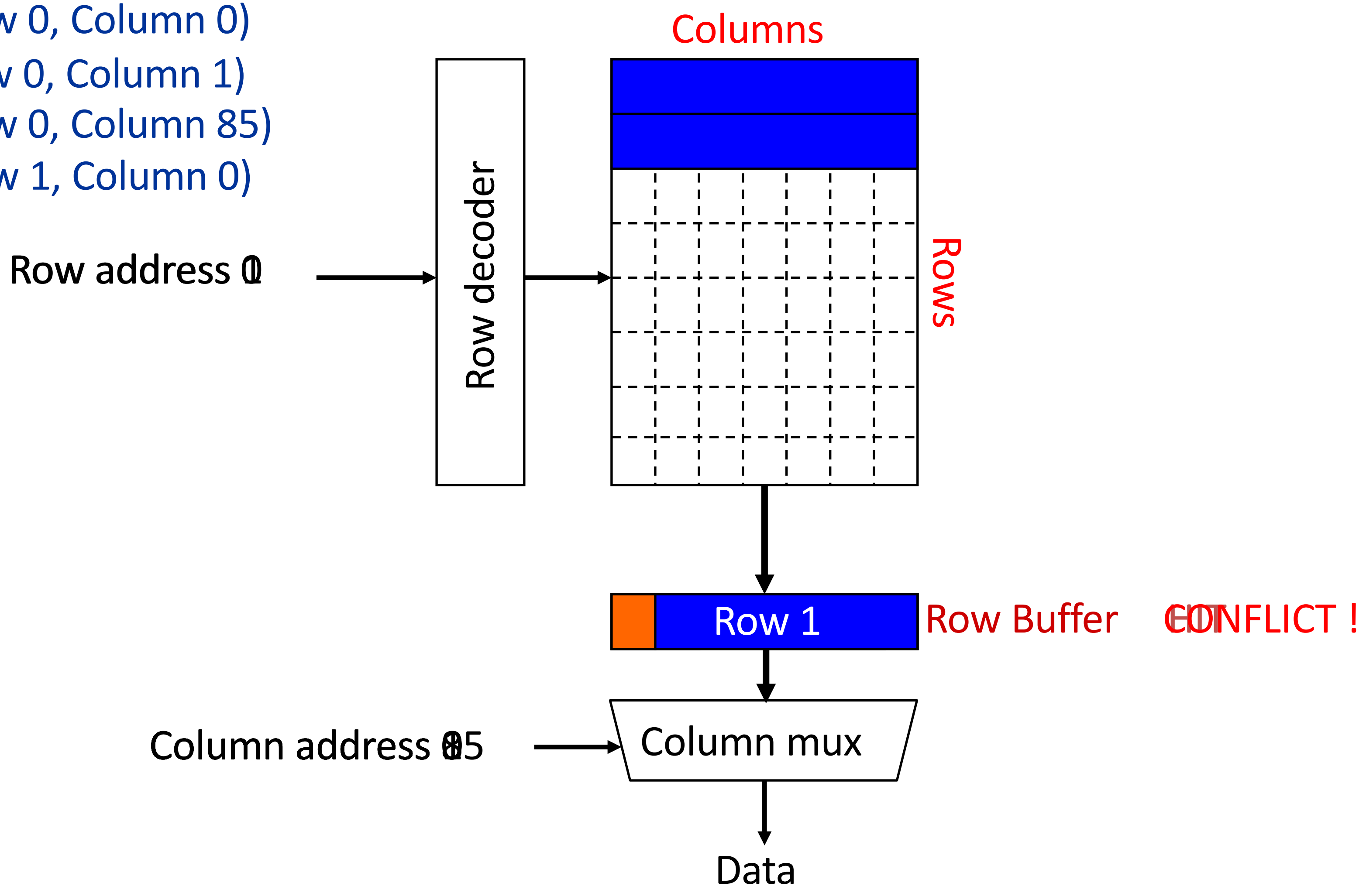
Cell= 1 Capacitor and 1 transistor

P.Nikolaou

Based on “Scalable Many-Core Memory Systems ” , Onur Mutlu, ACACES 2013

DRAM array Organization

Access Address:
(Row 0, Column 0)
(Row 0, Column 1)
(Row 0, Column 85)
(Row 1, Column 0)



P.Nikolaou

DRAM vs SRAM

	DRAM	SRAM
Capacity	✓ 1 capacitor + 1 transistor per bit	✗ 6 transistors per bit
Performance	✗ Refresh + probably larger transistors	✓ No Refresh +probably smaller transistors
Total Power	✗ Consumes power even when not used (periodic refresh)	✓ Access power
Cost	✓	✗
Reliability	✗ P.Nikolaou	✓

Memory Protection

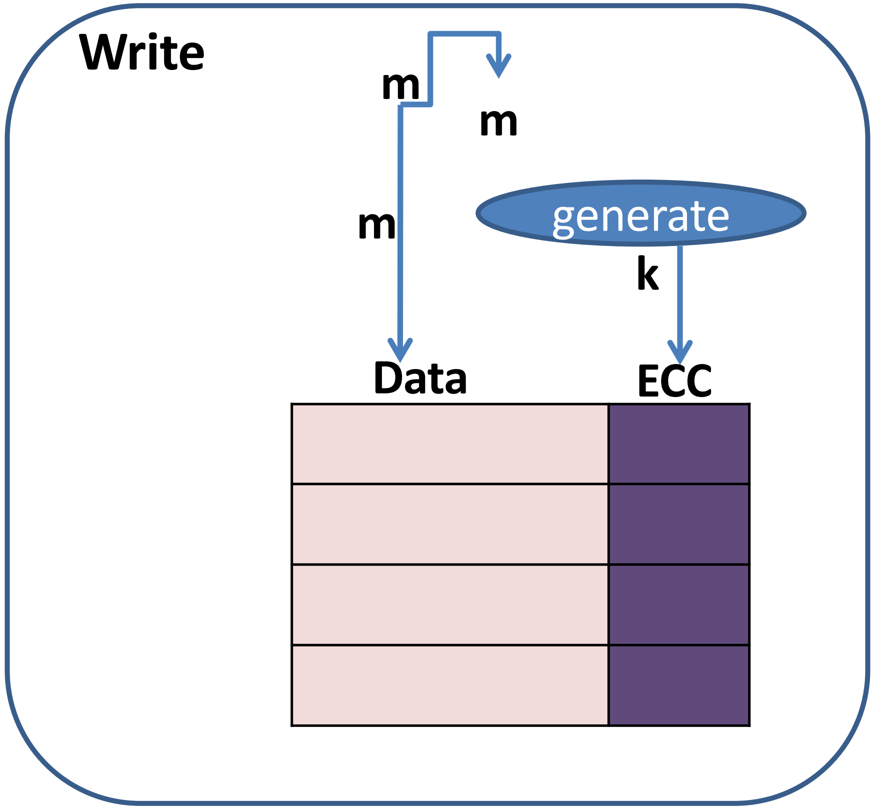
- **Data redundancy**
  - Error Detection/Correction Codes (EDC/ECC) [Hamming 1950, Hsiao 1970]
  - Applies in DRAM and cache
  - In DRAM, extra memory chips for ECC protection
  - Found in many variations
    - SEC-DED: Single Error Correction-Double Error Detection [Hamming 1950, Hsiao 1970]
    - DEC-TED: Double Error Correction-Triple Error Detection
    - ChipkillDC: corrects all errors that appear in a single memory chip and detects all errors that appear in two memory chips [AMD 2010, AMD 2014]
    - ChipkillSC: corrects all errors that appear in a single memory chip but cannot detect all errors that appear in two memory chips [AMD 2010, AMD 2014]

Protecting data from errors

How it works:

Write:

- Generate ECC bits(k) from data bits (m)
- Store data and ECC bits in the array

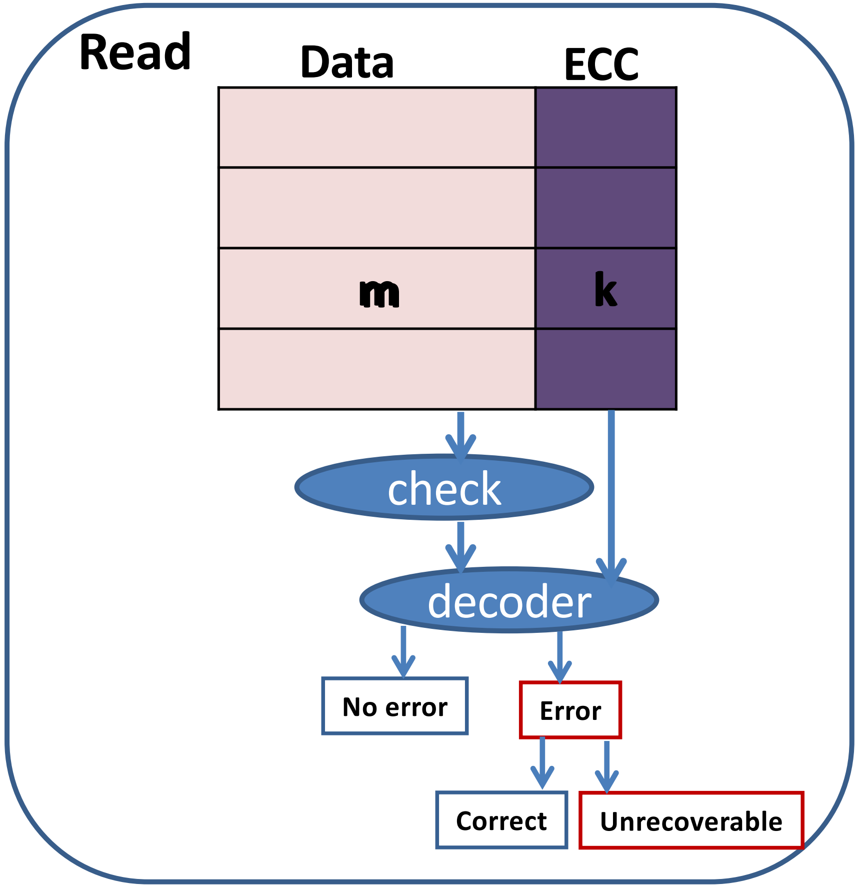


Protecting data from errors

How it works:

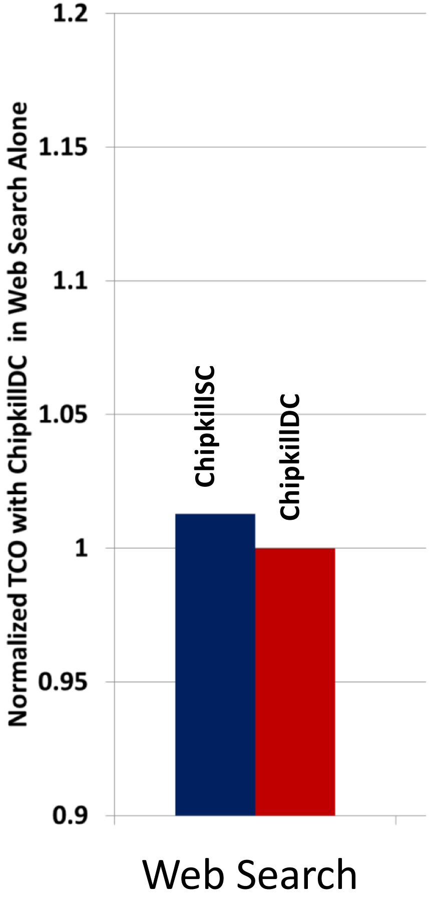
Read:

- Read data bits (m) and ECC bits (k) from the array
- Perform error checking
- The decoder indicates:
  - No error
  - Error:
    - Correctable
    - Uncorrectable



ChipkillSC Vs ChipkillDC

	ChipkillSC	ChipkillDC
Reliability	✓ Cannot detect all the errors in 2 devices Corrects all the errors in 1 device	✓✓ Detect all the errors in 2 devices Corrects all the errors in 1 device
Bandwidth	✓ Access one DIMM	✗ Access two DIMMs
Latency	✗ Codeword in two bursts	✓ Codeword in one burst
Power	✓ Access one DIMM	✗ Access two DIMMs


Normalized TCO with ChipkillDC in Web Search Alone
ChipkillSC
ChipkillDC
Web Search