Green Computing: Datacentres

Simin Nadjm-Tehrani
Department of Computer and Information Science (IDA)
Linköping University
Sweden

Many thanks to **Jordi Cucurull**
For earlier versions of this course material

This lecture

- Datacentre overview & trends
- Datacentre design
- Efficiency metrics
- Energy-proportional datacentres
- Initiatives

Recap from last lecture

- Basic power/energy related terms
- Useful notions to discuss energy use within ICT components
  - Energy-proportional computing
- Generic approaches for reducing energy
  - e.g. load consolidation (in time or space)

This lecture

- **Datacentre overview & trends**
- Datacentre design
- Efficiency metrics
- Energy-proportional datacentres
- Initiatives

Datacentres

“Datacenters are buildings where multiple servers and communication gear are co-located because of their common environmental requirements and for physical security needs, and ease of maintenance.”

Barroso and Hölzle

Luiz André Barroso and Urs Hölzle. The Datacenter as a Computer - An Introduction to the Design of Warehouse-Scale Machines. Morgan & Claypool Publishers, 2009 (120 pages).

Note: a 2013 2^{nd} Edition of this book (156 pages) is also available with more Google-oriented updates. Most of the basic material in the 2009 version is still worthwhile to study. The 2019 3^{rd} edition is at 189 pages!

Datacentre energy consideration

"Data centres total energy consumption in 2012 was about 270 TWh, which was about 2% of the global electricity consumption." (1)

Forbes: In 2017, global datacentres used 416 TWh, which was 3% of total electricity, and will double every 4 years. (2)

(1) W. Van Heddeghem, et al. Trends in worldwide ICT electricity consumption from 2007 to 2012, Computer Communications, Volume 50, Elsevier, 2014.

(2) https://www.forbes.com/sites/forbestechcouncil/2017/12/15/why-energy-is-a-big-and-rapidly-growing-problem-for-data-centers

Viral cat videos warming the planet?

- Not only energy consumption but also GHG emissions relevant

- https://www.theguardian.com/environment/2015/sep/25/server-data-centre-emissions-air-travel-web-google-facebook-greenhouse-gas

- GeSI 2015 report:

"Now we find that ICT can finally decouple economic growth from emissions growth" ☺

GeSI: Global e-sustainability initiative

Which data as a source of estimates?

- Green peace clicking green report (2017)
  - has mixed messages of what is getting better and what is getting worse
  - “IT sector is already estimated to consume approximately 7% of global electricity”

- Ericsson sustainability study (2010-2015) published in 2018:
  - “shows that the ICT and E&M sectors have turned their previously growing footprints into shrinking ones despite a continuous increase in subscriptions and data traffic”
  - Discusses what to include/exclude

https://www.mdpi.com/2071-1050/10/9/3027

GeSI 2019 report: SDGs and 2030 agenda

"It is estimated that emissions abated in 2030 as a result of greater adoption of these use cases will be equivalent to nearly seven times the size of the growth in the total ICT sector emissions between now and 2030."

"Digital technologies can also be deployed in ways that counter the Goals: fuelling consumption; hardening the digital divide; creating dislocation in the labour markets; and consolidating power of the few over the many. Enhancing the positive impact needs to go hand in hand with minimising any negative impacts."

https://gesi.org/research/gesi-digital-with-purpose-full-report

But...

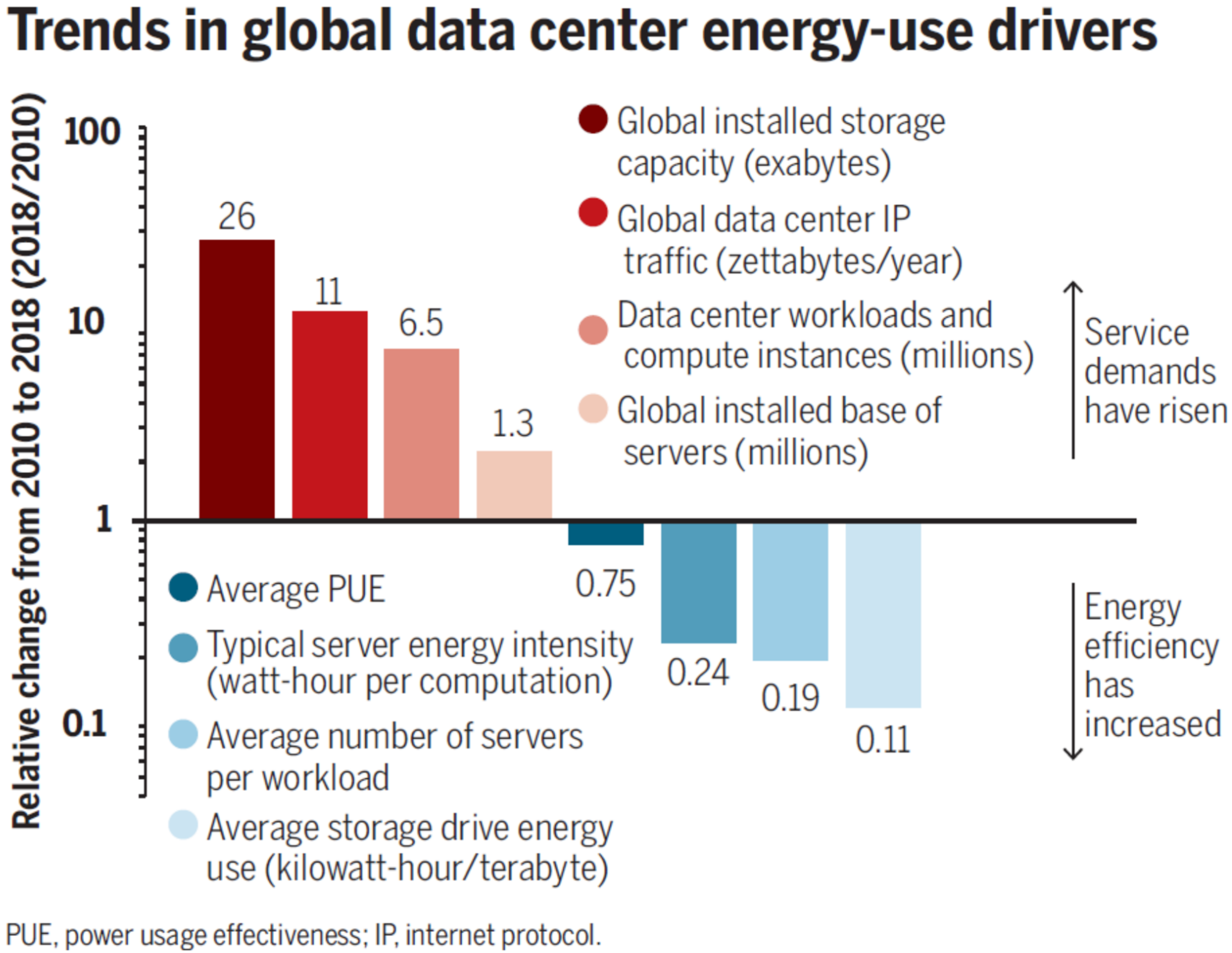
- Mentions seven ICT technologies as vehicles for reaching103 of the 169 SDG targets through ICT

- Full connectivity
- Fast Internet (5G)
- Cloud
- IoT
- Cognitive (AI/ML/analytics)
- Digital reality (AR/VR)

and…

- Blockchains!

Compensating trend



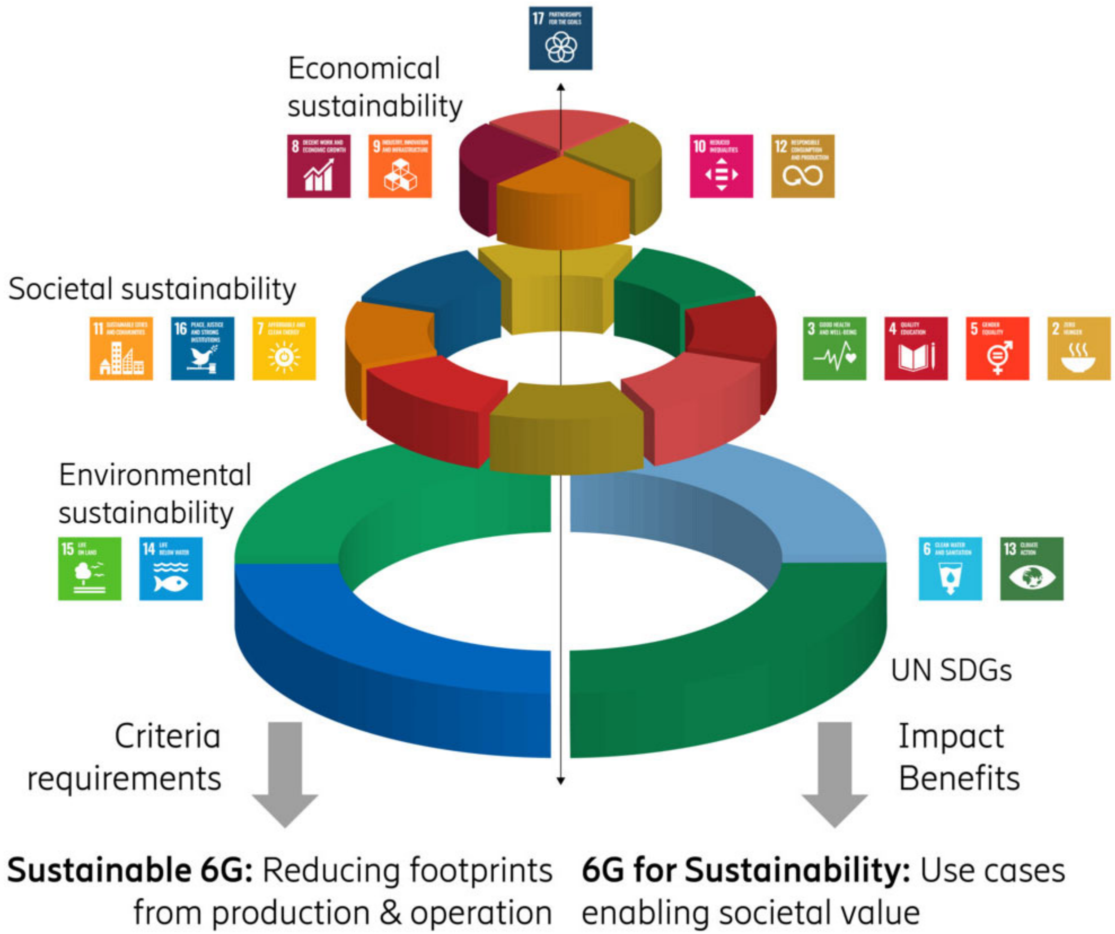
Masanet et al. February 2020, DOI: 10.1126/science.aba3758

We are now in the age of AI…

- Do we expect that the emissions and electricity use today is growing slower than the growth in rate of data transmitted, processed, stored?

- Recall the revolutionary impact of AI from the last lecture!

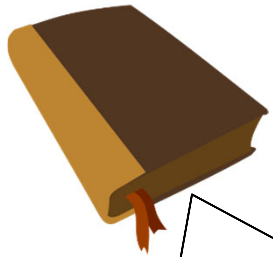
6G in horizon



From https://www.ericsson.com/en/blog/2023/8/6g-sustainability-first

This lecture

- Datacentre overview & trends
- **Datacentre design**
- Efficiency metrics
- Energy-proportional datacentres
- Initiatives

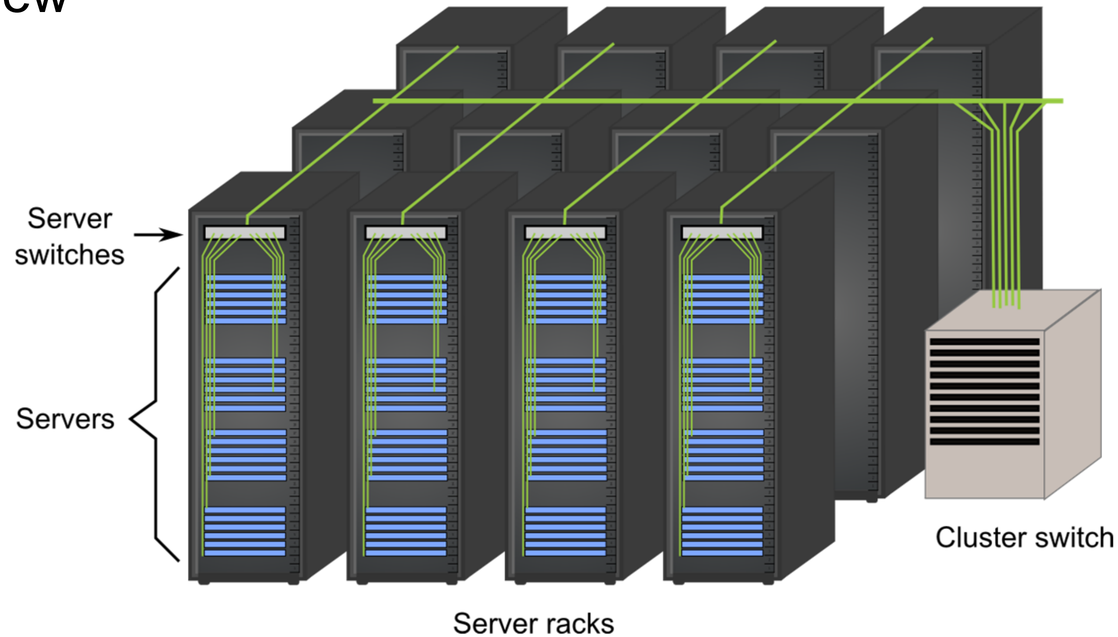


Chapter 4 and 5 of the 1st Edition of the e-book on the course literature (2009), and updates from later versions

Let's look into a Datacentre!

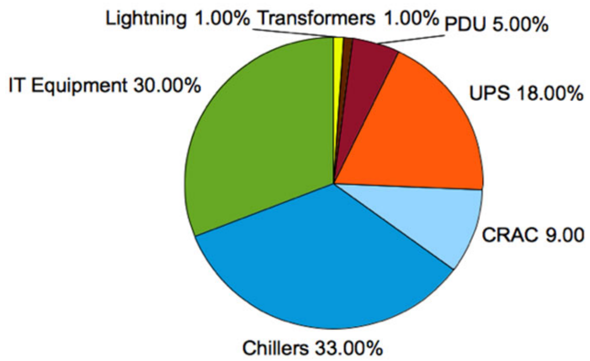
IT Architectural view

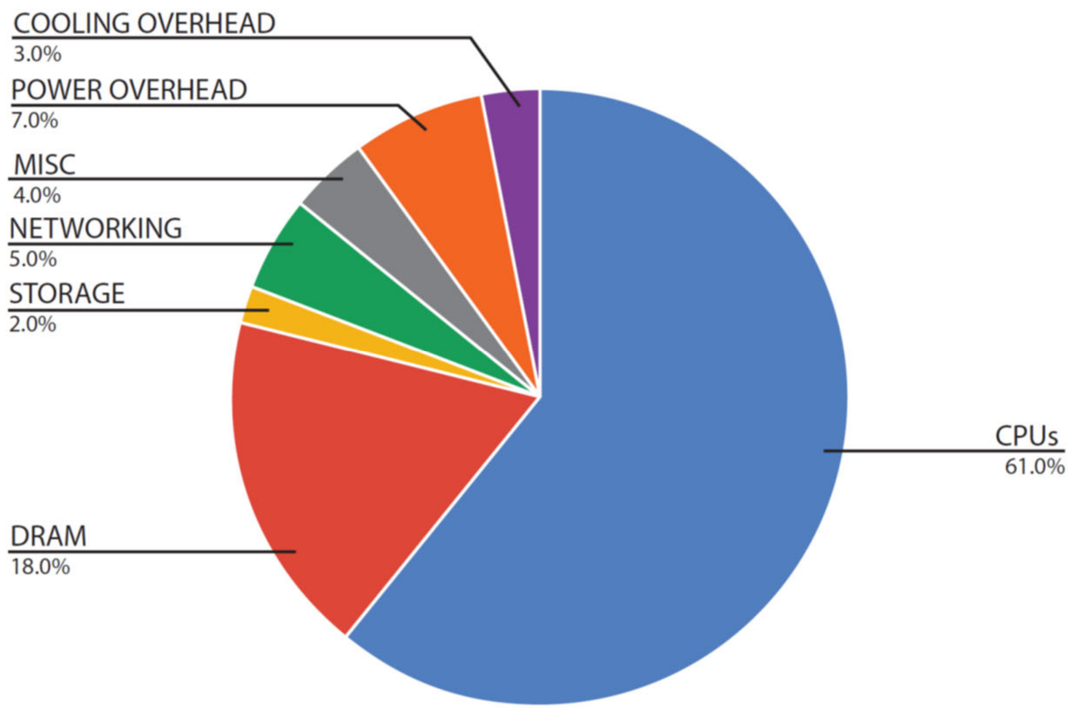
Server



Check out virtual tours on the Internet!
https://www.youtube.com/watch?v=80aK2_iwMOs

Energy use at a datacentre

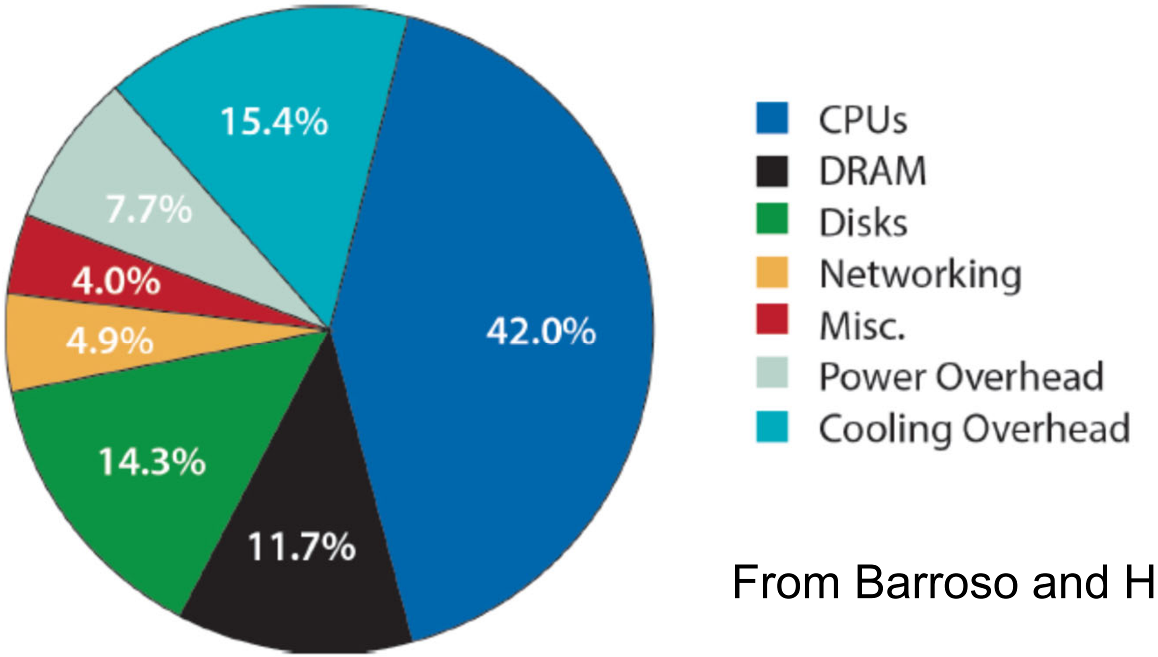




Datacentre overhead (Left: Data from one Google facility 2007) (Right: a typical data centre 2012)

From Barroso and Hölzle, 2009 and 2019

Server energy consumption

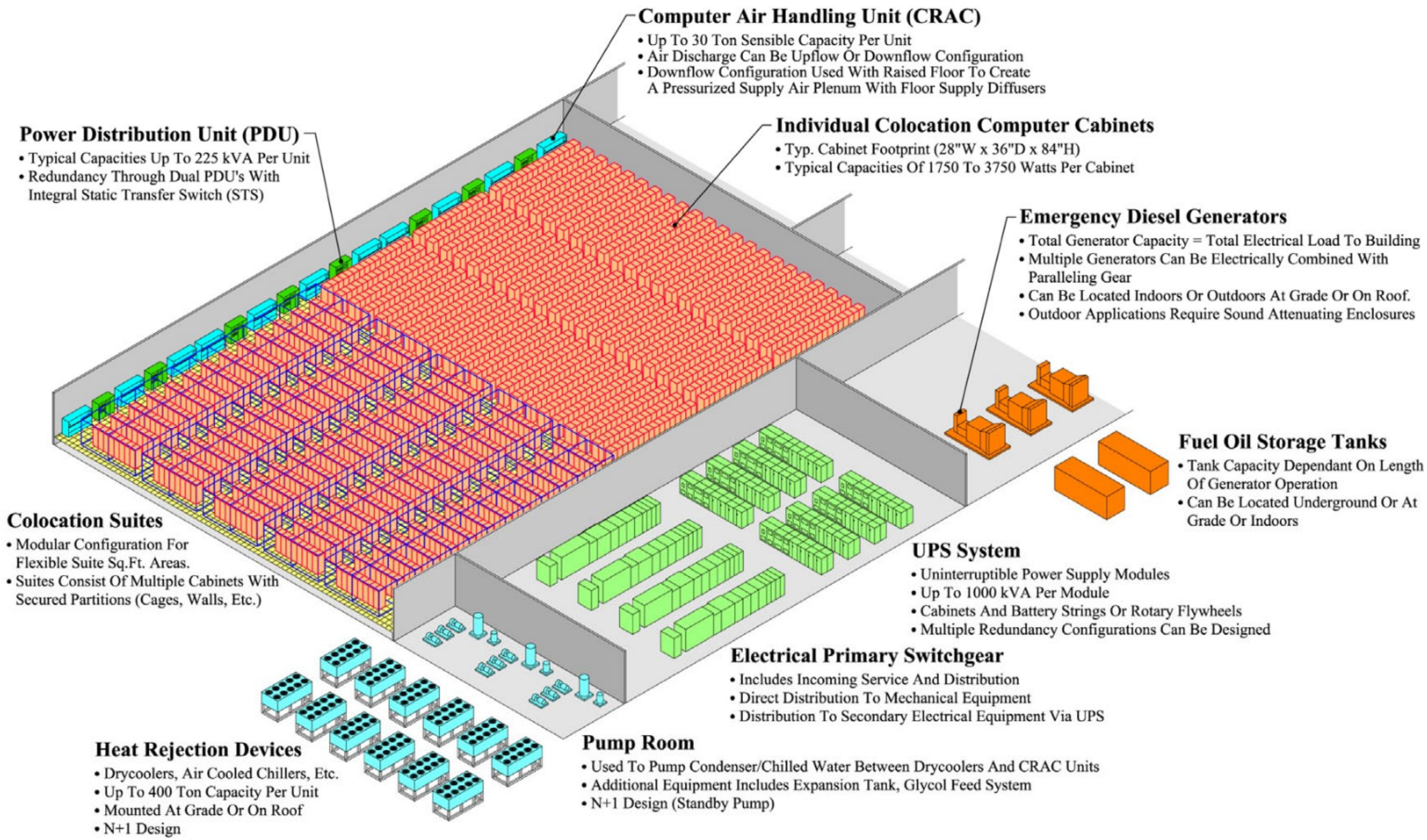


From Barroso and Hölzle (2013)

- 2024: Direct Liquid Cooling (DLC) promises gains in heat reduction and reduced water consumption

https://www.data4group.com/en/csr/liquid-cooling-in-data-centers-a-revolution-in-energy-efficiency/

Classic DC design: Power and cooling



Graphic provided courtesy of DLB Associates

Power system components

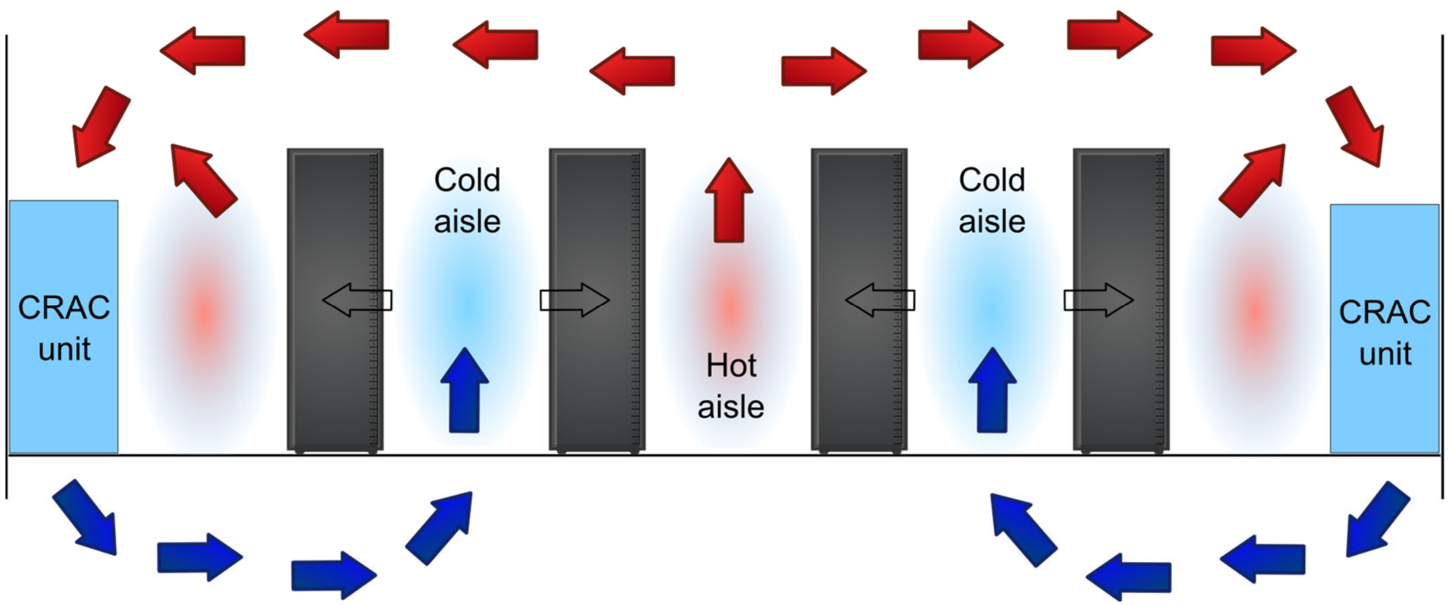
- Primary switchgear
  - Breakers for protection
  - Transformers (10-20kV to 110-600V)
- Diesel generators
  - Switched on in case of utility power failure
- Uninterruptible Power Supply (UPS)
  - Batteries for short term energy provision
  - Functionality
    - Switch energy source between mains power (from utility) and power from Diesel generators
    - Sustain system power with batteries during mains power failure
    - Power feed conditioning

Power system components

- Power Distribution Units (PDUs)
  - Break a higher voltage line into several circuits
    - 200-480V line to many 110-230V circuits that feed the servers
    - A ground short (in server or power supply) will only break one circuit
  - Distribute energy to each rack
  - Provide redundancy (A-side, B-side) so that with a power supply failure fast switching can take place

Standard cooling system

- Computer Room Air Conditioning (CRAC)
  - Blows cold air under the floor plenum
  - Cold air moves to front of server racks (cold aisle)
  - Cold air flows through server racks
  - Warm air is expelled in the back (warm aisle)



Cooling: operational range

- Datacentre cannot operate without cooling
- CRAC units cool the room's air
  - Liquid coolant is pumped from chillers or cooling towers
  - Coils are kept cool (12-14 ºC) with liquid coolant
  - Warm air is pushed through the coils by fans
  - Cold air (16-20 ºC) is moved to the floor plenum
- Air reaches the servers at 18-22 ºC

Free cooling

- Much more efficient than chillers
- Cool the coolant to much lower temperatures before reaching the chiller

Cooling tower

Glycol-based radiator

Fans to push air from outside

Redundancy and reliability

- Redundancy for fault tolerance and during maintenance
  - Applied to both power and cooling systems

- Configurations: N+1, N+2, 2N

Tier Classification (ANSI/TIA 942)

Type	Availability	Description
Tier I	99.67%	Single path for power & cooling. No redundancy (N)
Tier II	99.74%	Single path for power & cooling. Redundancy (N+1)
Tier III	99.98%	Multiple paths for power & cooling (only one active). Concurrently maintainable. Redundancy (N+1)
Tier IV	99.995%	Multiple active paths for power & cooling. Redundancy in both paths (min (N+1))

W.Pitt Turner IV, J.H.Seader, K.G.Brill. Tier classifications define site infrastructure performance, Uptime Institute, White Paper

Container-based datacentres

Highly efficient cooling, NYT Feb. 2016

http://news.microsoft.com/features/microsoft-research-project-puts-cloud-in-ocean-for-the-first-time/#sm.0000q5ts4lqgfez110wem1gb0ig5o

This lecture

- Datacentre overview & trends
- Datacentre design
- **Efficiency metrics**
- Energy-proportional datacentres
- Initiatives

Sources of efficiency losses

- Power systems

Element	Losses
Transformers	0.50%
UPSs	7-12%
Highly efficient UPSs	3%
Low-voltage power (110-220V) cables	1-3%

- Cooling
  - Fans that move cool and warm air
  - Mix of cool and warm air during long paths
  - Too low temperature selection
    - 25-27 ºC better than targeted 20 ºC

Energy efficiency metrics

- Power Usage Effectiveness (PUE)
  - Proposed by The Green Grid association
  - Efficiency of the IT support infrastructure

$$PUE = \frac{Total\ facility\ power}{Total\ IT\ equipment\ power}$$

Historic data: PUEs between 1.5 and 2.0

Google, Microsoft, Apple have reported PUEs around 1.1X

C. Beladi. Green Grid datacenter power efficiency metrics: PUE and DCiE. White paper. 2008

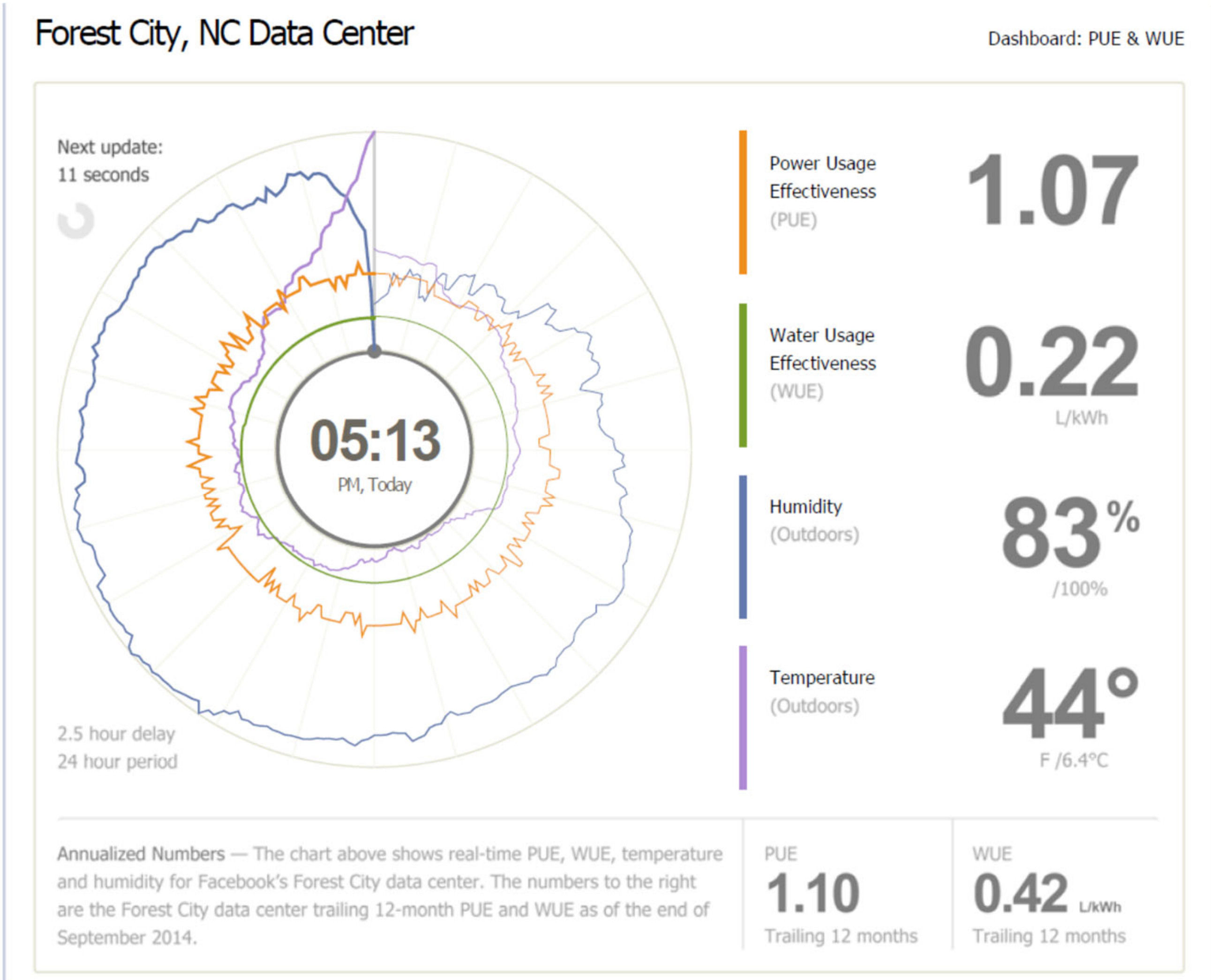
Recent data: https://www.google.com/about/datacenters/efficiency/

Location matters…

- One quarter of Singapore data centers have PUE higher than 3, Japan and Australia claim average PUE of 2.2 and 2.25 respectively (APAC datacenter survey, April 2013)
- Facebook Luleå: 1.08 (Computer Sweden, 7 Feb 2013)

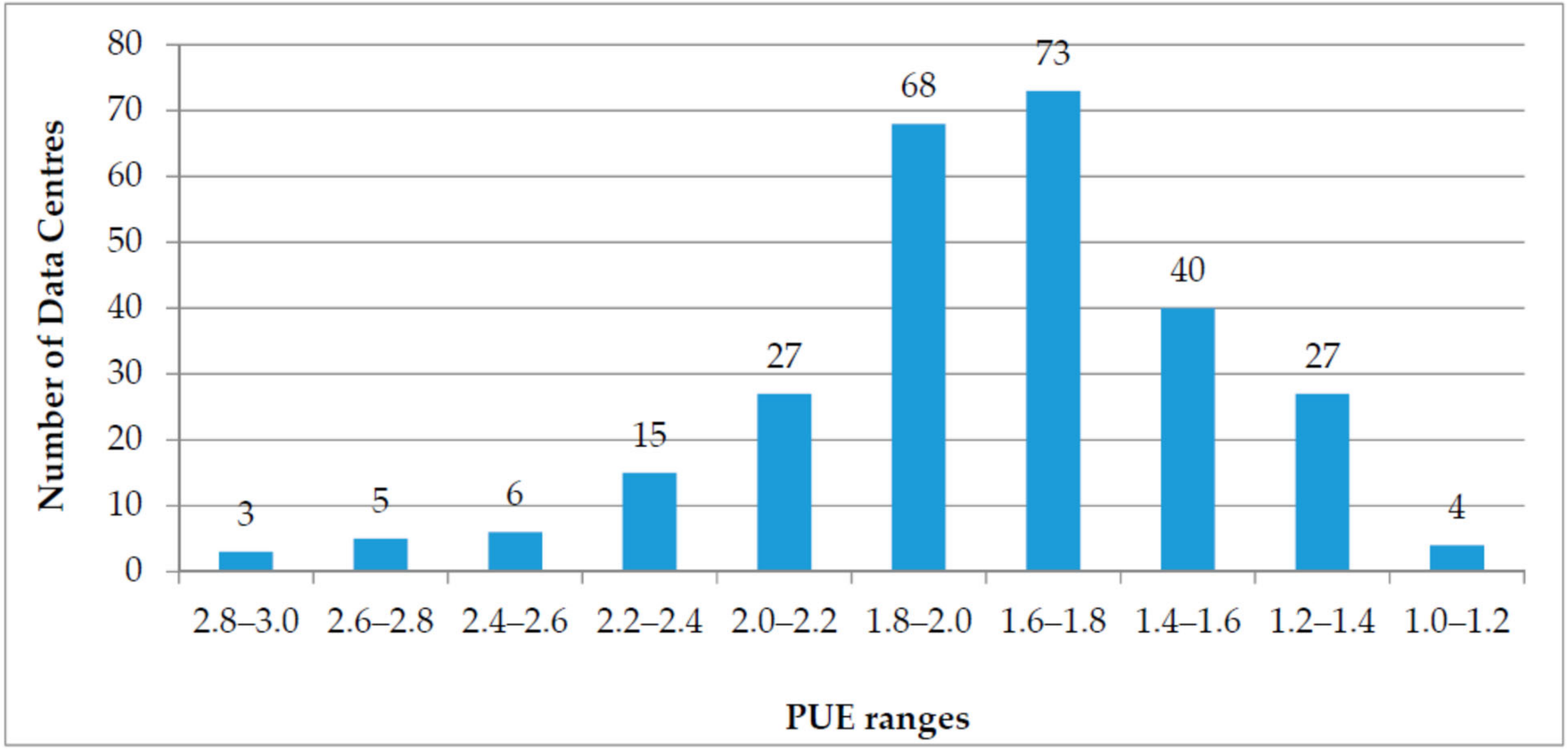
Image: https://www.facebook.com/LuleaDataCenter

Facebook live dashboard



https://www.facebook.com/ForestCityDataCenter?sk=app_288655784601722&app_data

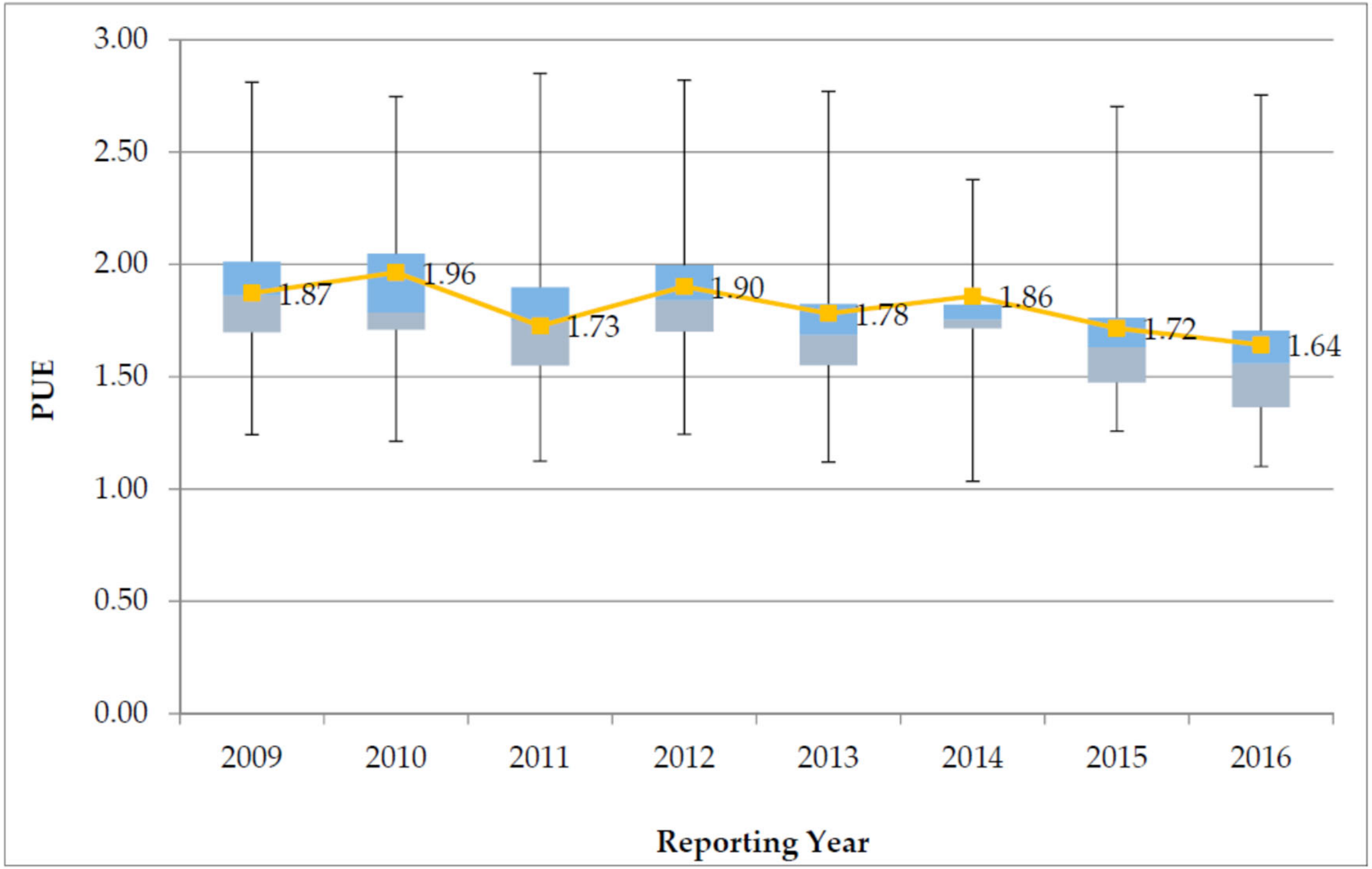
What about Europe?



Data from the 2017 EU study performed by JRC on 268 datacenters in Europe

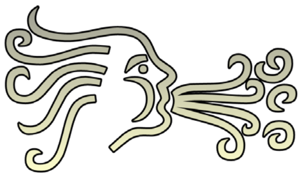
http://www.mdpi.com/1996-1073/10/10/1470/htm

Big reductions? Not really…



Energy efficiency metrics

- PUE has to be used carefully
  - <u>Example</u>: Server fans in IT-PAC (Pre-Assembled Container) Microsoft datacentre container module



$$\downarrow \frac{\textit{Facility power}}{\boxed{\textit{IT equipment power}}}$$

IT-PAC airflow management is improved ⇒ Server fans are no longer required ⇒ PUE increases

Overall energy consumption reduced despite PUE increase

- Other aspects must be taken into account

http://www.datacenterknowledge.com/archives/2011/01/31/microsoft-eliminates-server-fans-despite-pue-hit/

Energy efficiency metrics

- Green Grid's Datacentre Performance Efficiency (DCPE)

$$DCPE = \frac{Amount\ of\ computational\ work}{Total\ energy\ used}$$

- Barroso and Hölzle propose:

DCPE = 1/PUE . 1/SPUE . Computation/Total energy to electronics

Computing efficiency

IT equipment efficiency

Building facilities efficiency

Energy efficiency metrics

- Server Power Usage Effectiveness (SPUE)
  - Efficiency of the IT infrastructure
    - Losses in power supply, voltage regulator modules, server fans

$$SPUE = \frac{\text{Total server input power}}{\text{Total useful power}}$$

    - Useful power
      - Consumption of electronics directly involved in computation
      - E.g. motherboard, disks, CPU, DRAM...

Servers have SPUEs between 1.6 and 1.8 (Barroso & Hölzle 2009)

State of the art servers should be less than 1.2!!!

At what load?

Energy efficiency metrics

- Total Power Usage Effectiveness (TPUE)
  - Show efficiency of electromechanical overheads.

$$TPUE = PUE \times SPUE$$

Example: 2.0 PUE
1.6 SPUE

$$TPUE = 2.0 \times 1.6 = 3.2$$

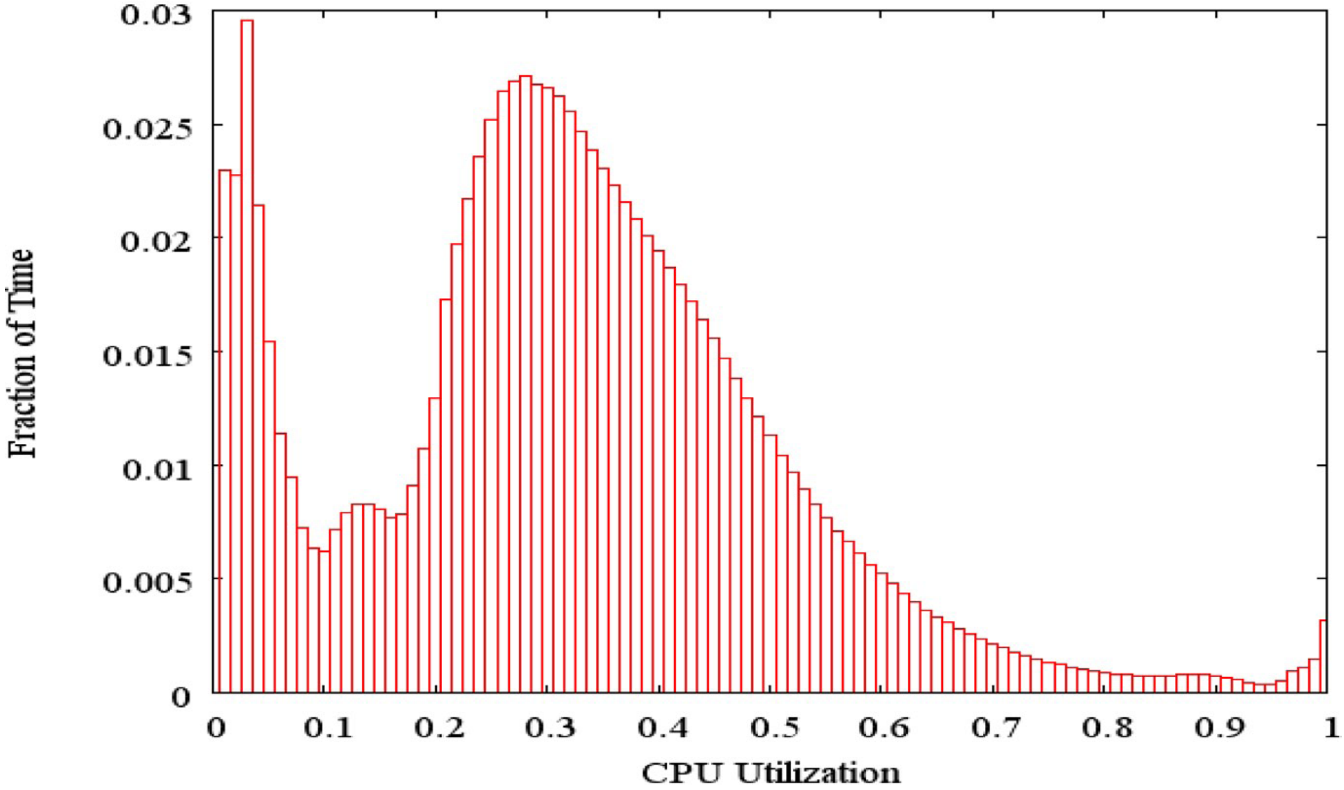
For each productive Watt
another 2.2 W consumed!!!

This lecture

- Datacentre overview & trends
- Datacentre design
- Efficiency metrics
- **Energy-proportional datacentres**
- Initiatives

Load and energy efficiency

- Most of the time spent in non energy-efficient load regions



Activity profile of 5000 Google servers in 6 months

Image from Barroso and Hölzle (2009)

Load and energy efficiency

- Almost no time in idle state

- Load spread out over all available servers
  - Optimised for availability and performance

Sustainable approach

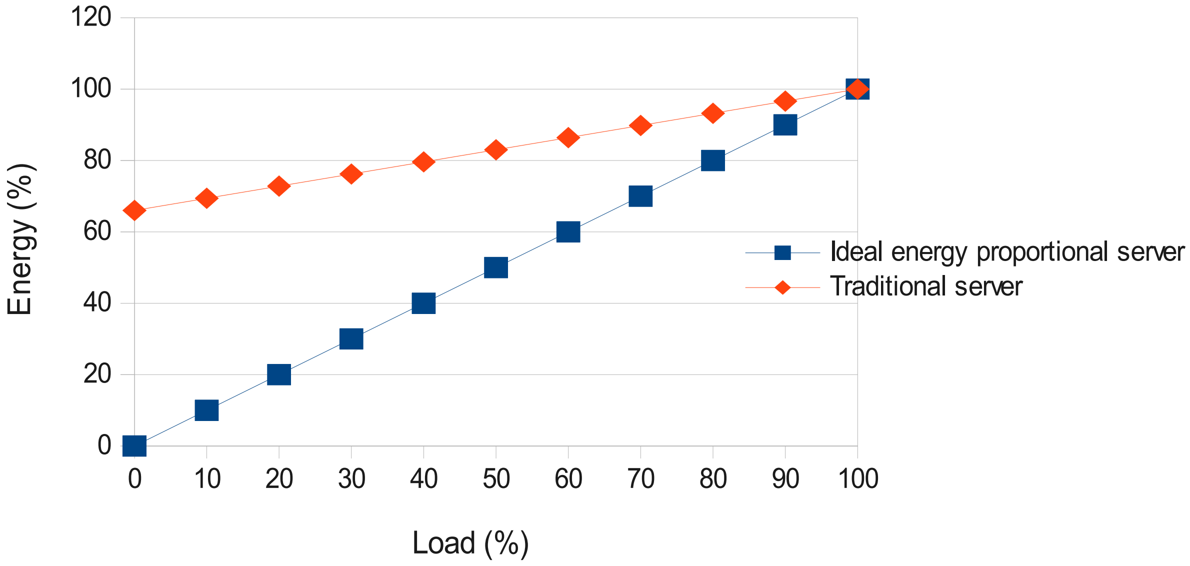
Concentrate load in a few servers
Switch off unused servers

But!!!

More software complexity
Energy spent for data distribution

Recall from lecture 2

- Consumption proportional to the load
  - Ideal linear function without constants



Traditional server data based on data from Barroso and Hölzle (2009)

Energy proportional computing

- Capacity to adapt consumption to load (the higher the better)

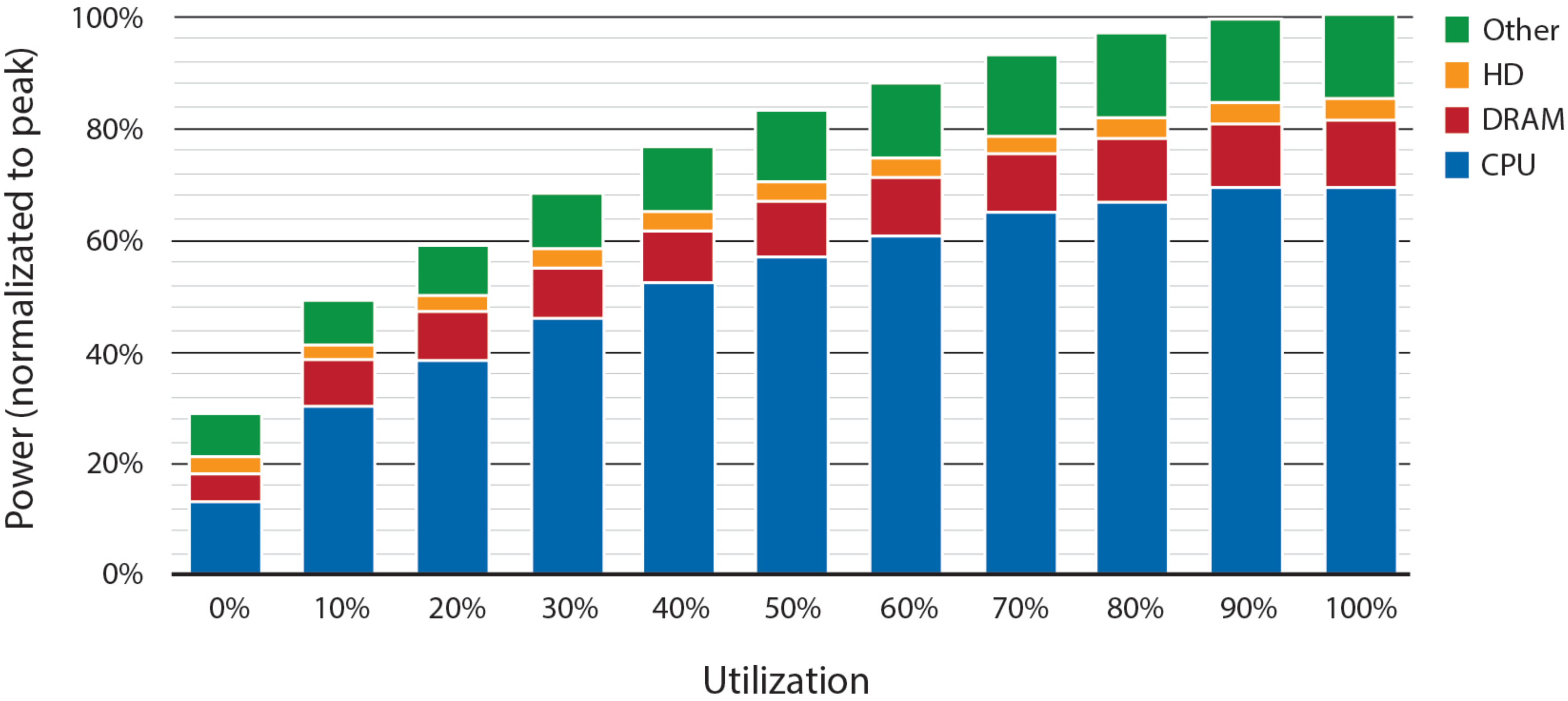
Device	Dynamic power
CPU	3.0x
Memory	2.0x
Disks	1.3x
Network switches	1.2x

More hardware improvements are required!!!

Meanwhile only option is to switch off hardware!!!

Energy spent: revisited

CPU a larger proportion of energy costs than that reported in 2009!



Barroso and Hölzle (2013)

This lecture

- Datacentre overview & trends
- Datacentre design
- Efficiency metrics
- Energy-proportional datacentres
- **Initiatives**

Industry sector specifications

- ENERGY STAR
  - Program Requirements for Computer Servers 2.0 (Dec 2013)
    - Includes active consumption
      - as opposed to just idle consumption
  - Requirements for Datacenter Storage 1.0 establishes maximum consumption of storage products
- 2014: Microsoft joined Facebook to lead the Open Compute Project on servers
- 2019: Intelligent cloud!
- Today's focus is more on building AI-friendly hardware, will that be energy-efficient too?

EU Code of Conduct for Datacentres

- European action to reduce consumption of datacentres
  - Best practices for datacentres operators
    - V3.0 Guide (Feb 2011)
  - Operators register and commit to their application

CODE OF CONDUCT
DATA CENTRES

- Roles of applicant
  - Participant
    - Operator of datacentre or equipment in it
    - Commitments
      - Annual reporting of energy consumption
      - Implementation of some of the best practices
  - Endorser
    - Supports the initiative and participants

https://e3p.jrc.ec.europa.eu/communities/data-centres-code-conduct
2024 report (75 pages)

To sum up

- Datacentres are a big part of ICT energy puzzle

- Big companies' competition and publicity has led to improvements in power efficiency
  - But we seem to have reached a "limit"
  - Public advocacy and regulations seem to play a role
  - Improvements always have a cost/benefit trade-off