Compute systems: architectures, infrastructures and energy efficiency

Séminaire Informatique distribuée et calcul haute performance

Gilles Sassatelli
sassatelli@lirmm.fr

+ colleagues: A. Gamatié, M. Robert, L. Torres

@Virtual,– July 1st, 2021

Horizon 2020
European Union funding
for Research & Innovation

Outline

Fundamentals: energy, CO_2 & the Joule Journey

Tiny introduction to supercomputer & processor architecture (power-wise)

Tradeoffs in multicore processor design

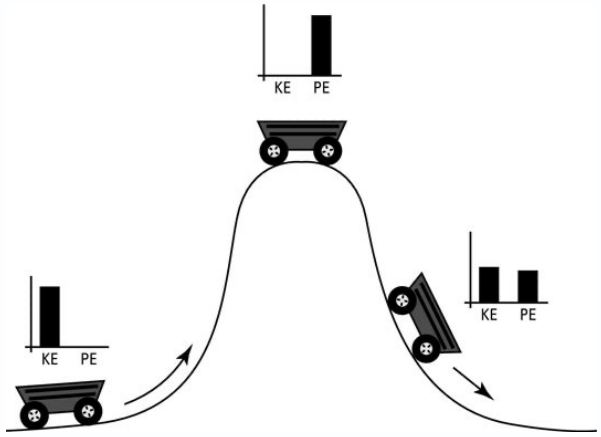
Playing with runtimes for energy efficiency

Levers at infrastructure-level: is carbon neutrality even achievable?

Energy & CO_2

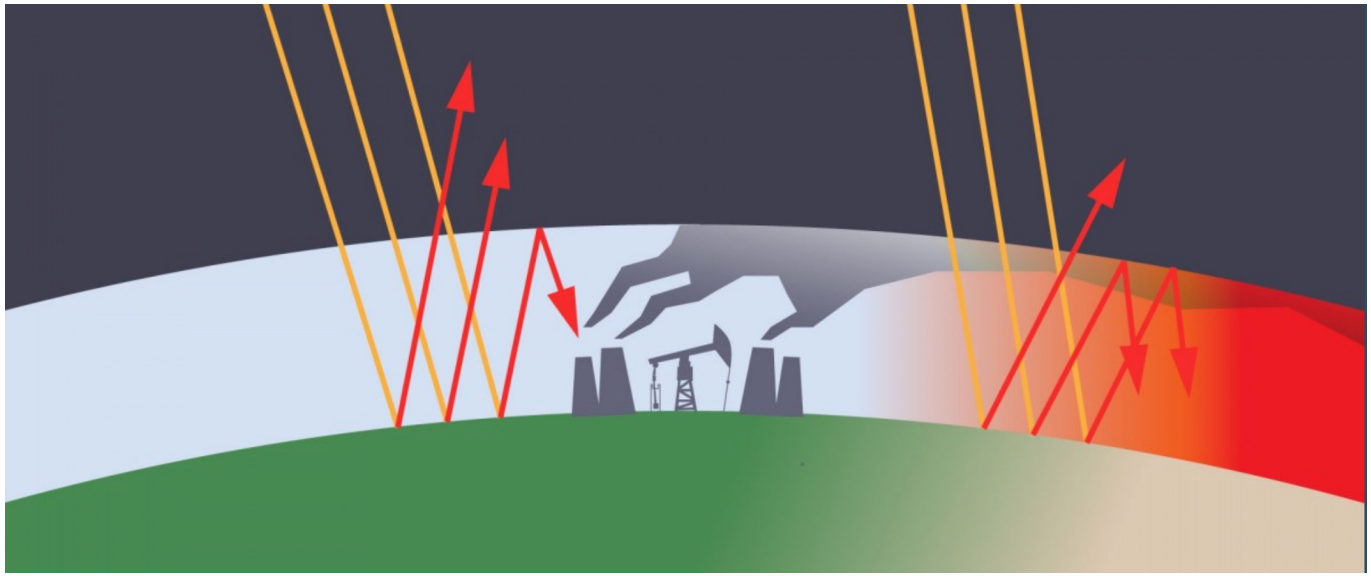
Energy ~ measure of ability to change state

- Conserved
- Storable: kinetic, potential, chemical
- Fossile fuels stores quite an amount (oil ~10kWh/l)



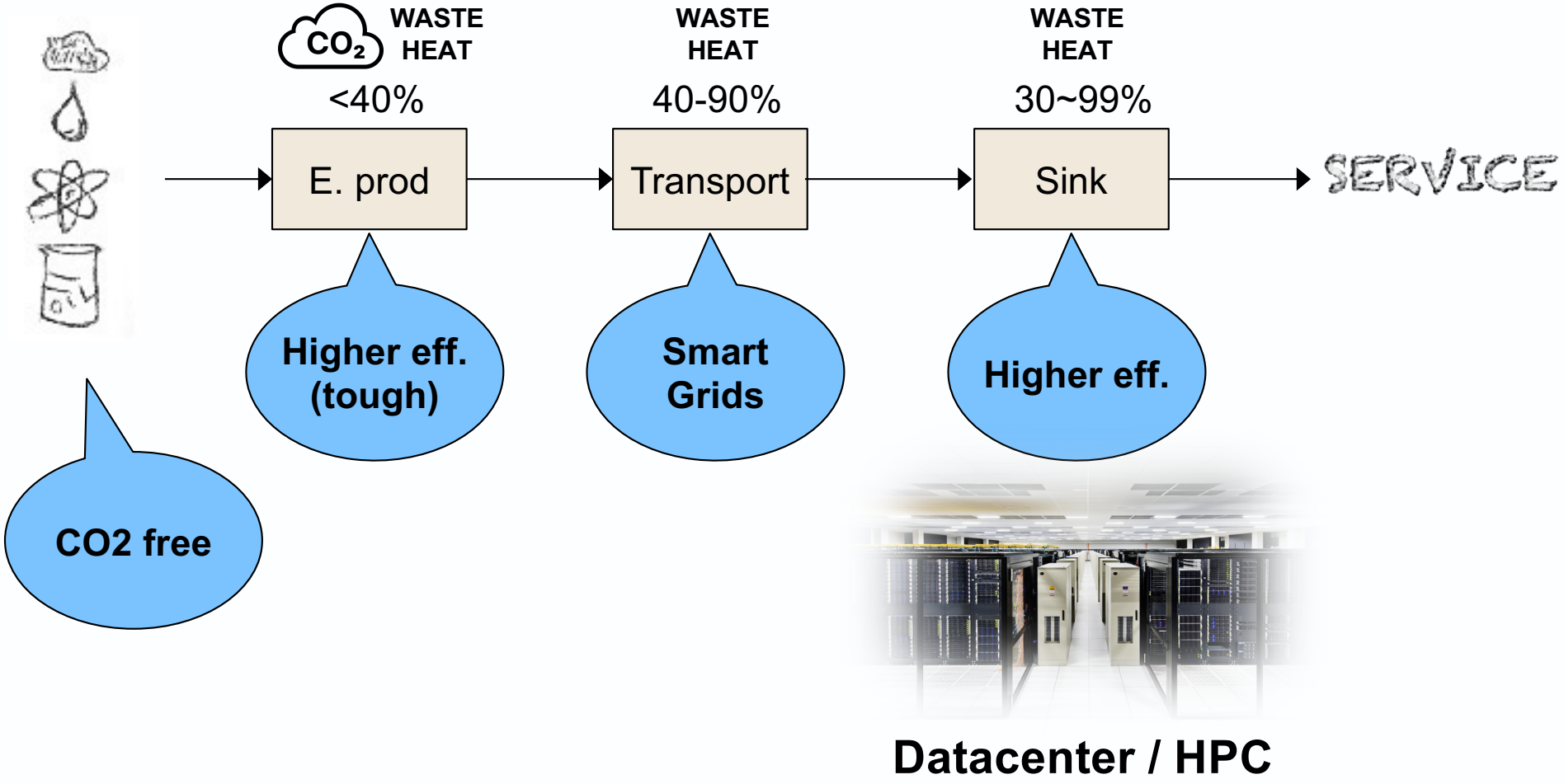
CO_2, the nasty byproduct (and not heat per se)

- Sun: 174.10^{15} W → ~1.5h collection power earth for 1 year
- Disturbingly stable as an oxyde: there to last
- Loves infrared produced by mildy warm bodies (earth)



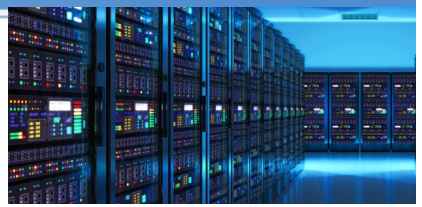
Energy, transformation, delivery and use

Energy efficiency, electricity delivery



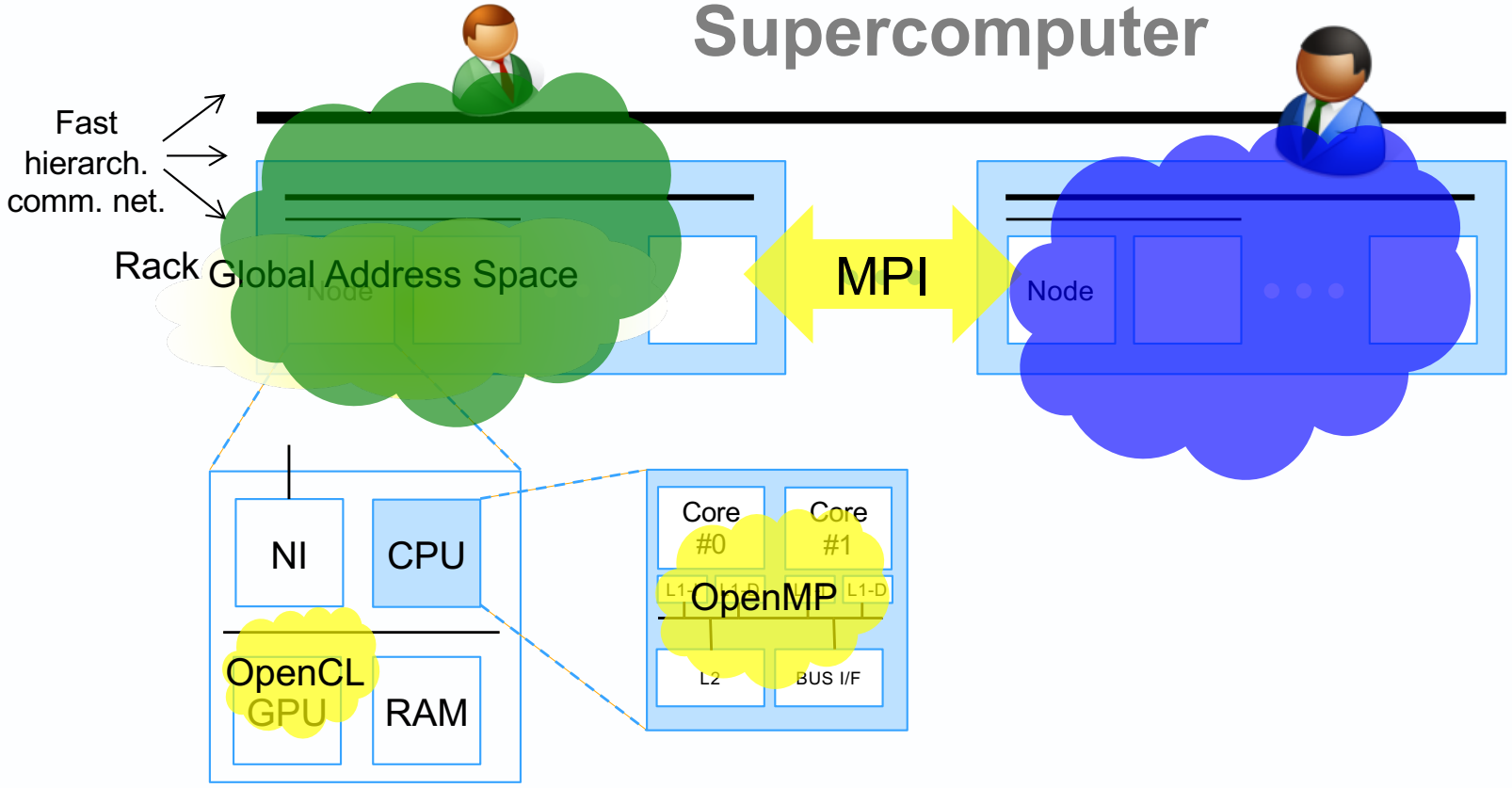
http://insideenergy.org/2015/11/06/lost-in-transmission-how-much-electricity-disappears-between-a-power-plant-and-your-plug/

What does a typical DC / Supercomputer look like?



Key concepts

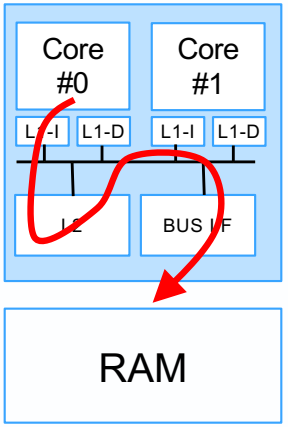
- Site with infrastructure equipment hosting servers, mass storage w. networking equipment
- Hardware installed on blades organised in racks in server rooms
- Supercomputer or Data centre? Same ingredients, different recipe!



Ooooooversimplification

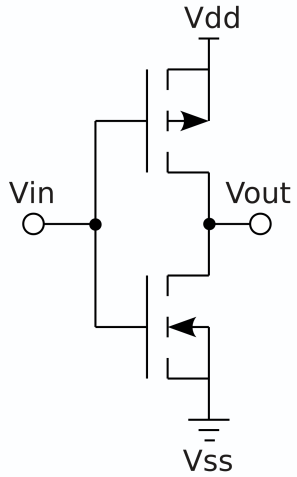
Let’s start with a (very) weak assumption

- Compute is performed by processors
- And from there mostly originates power consumption



Von Neumann legacy

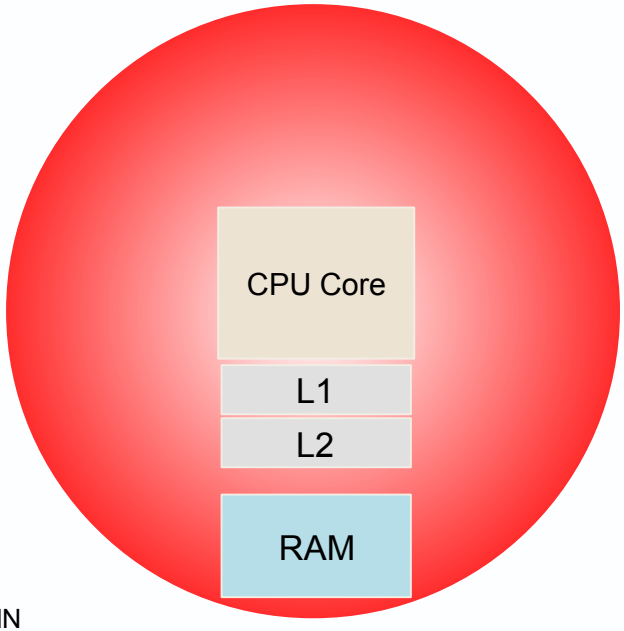
- Data travels back and forth with memory
- Cached in .. Caches
- All CMOS (Complementary Metal Oxide Semiconductor) logic
  - > 10^7 transistors



Understanding where power goes

Compute or **communicate**? *On-chip figures for 22nm CMOS*

- **Communicate**:
  - Moving 1bit of information: 1pJ/mm
  - @1GHz: 1pJ/mm x 10^9 s^{-1} = 1mW/mm
  - On a 64 bit bus, 64mW/mm
- **Compute**:
  - Toggling 1bit: 1aJ =10^{-6}pJ

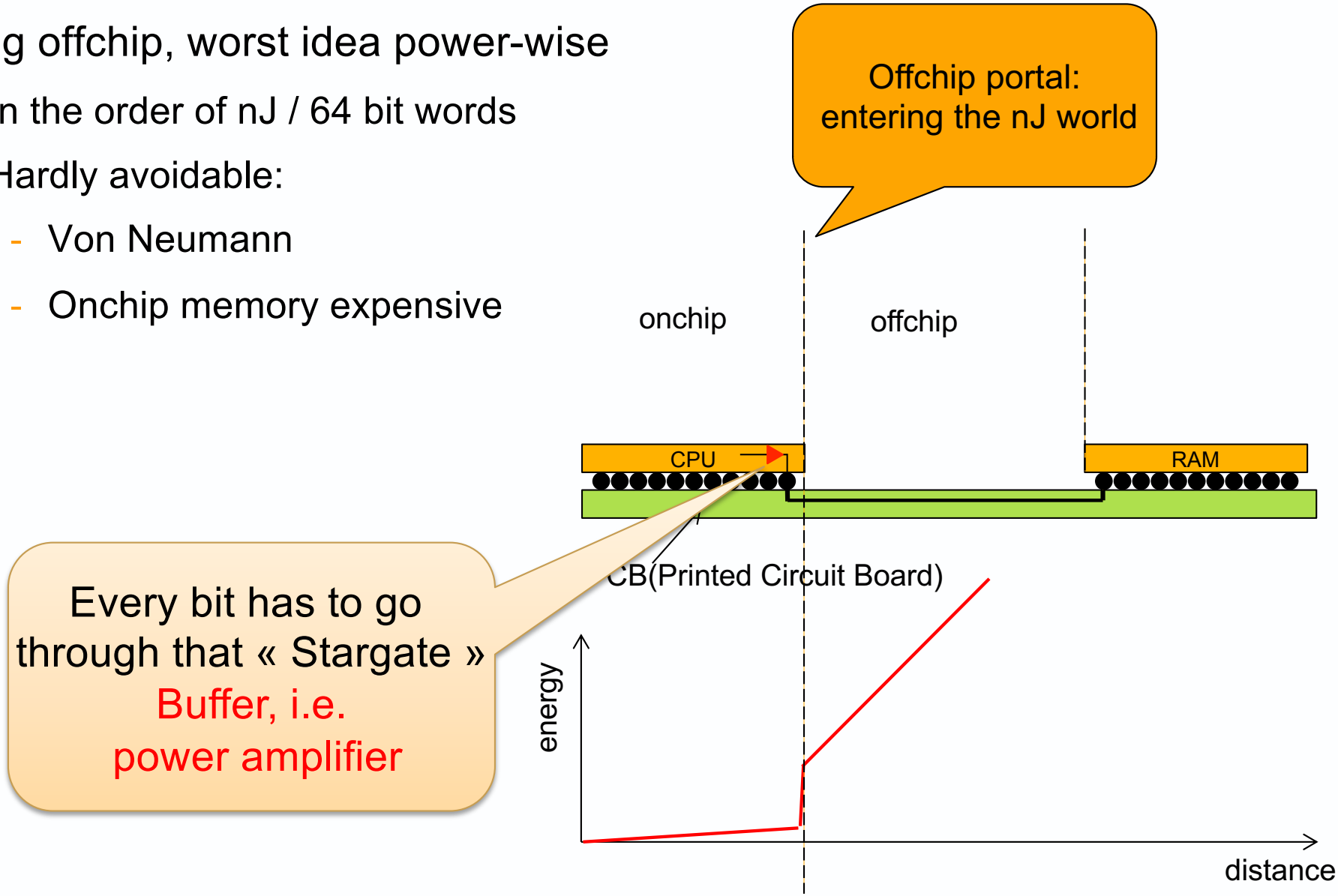


Source: H. Fanet, CEA LETI; A. Cappy, IEMN

Understanding where power goes

Going offchip, worst idea power-wise

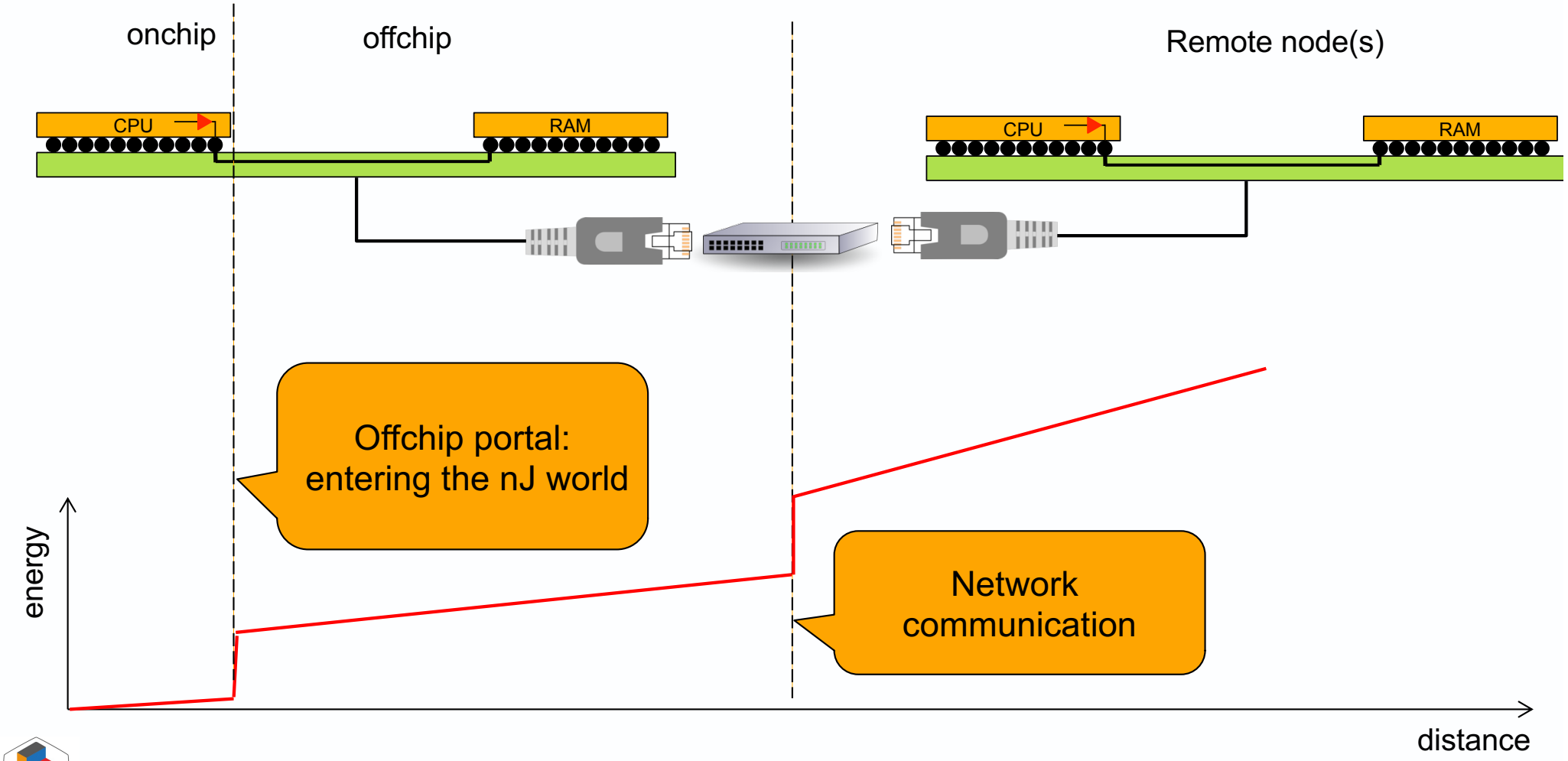
- In the order of nJ / 64 bit words
- Hardly avoidable:
  - Von Neumann
  - Onchip memory expensive



Understanding where power goes

Worse for computer clusters i.e. Supercomputers

- Better increasing node compute density (less nodes, less remote comms)



Drafting design guidelines

Data should not travel much!

- Undergo as much processing as possible *in place*
  - *Pack up lots of cores per node*
  - *Data fetched from DRAM should not go back and forth*
- In the most possible energy efficient way

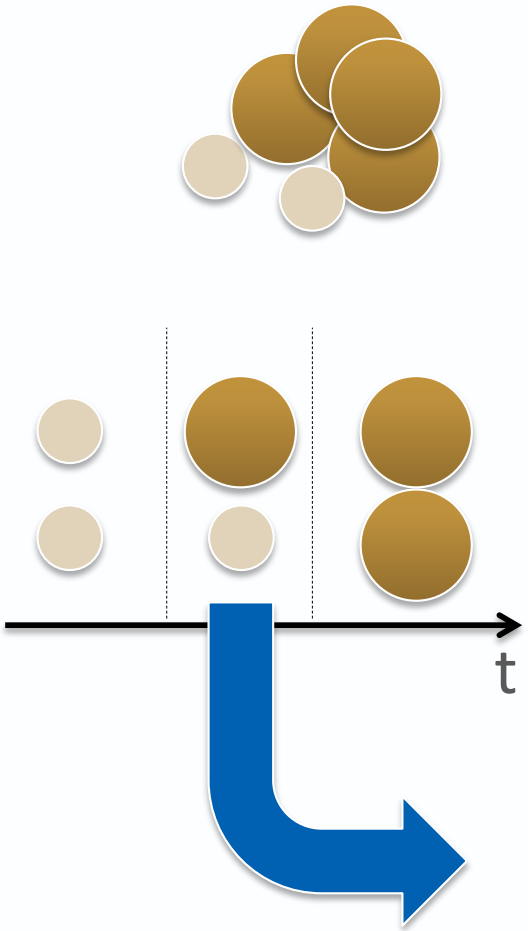
Operation	Energy
Addition of data (fixed point)	1x
Access data (onchip cache)	60x
Access data (offchip RAM)	3500x

Pedram et al , IEEE D&T 2016

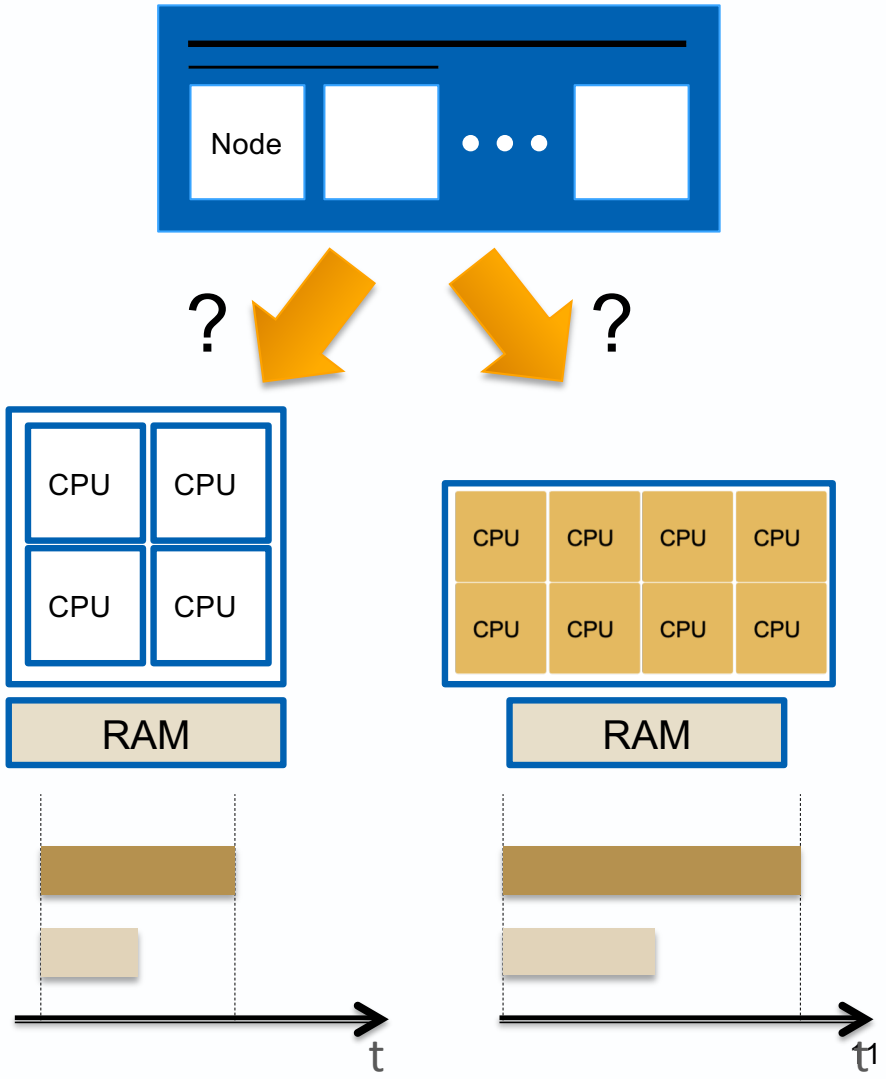
Approach

Rationale: the closer the tasks, the better: favor locality i.e. *in-place*

- Compute jobs = Tasks(t)



- Cluster

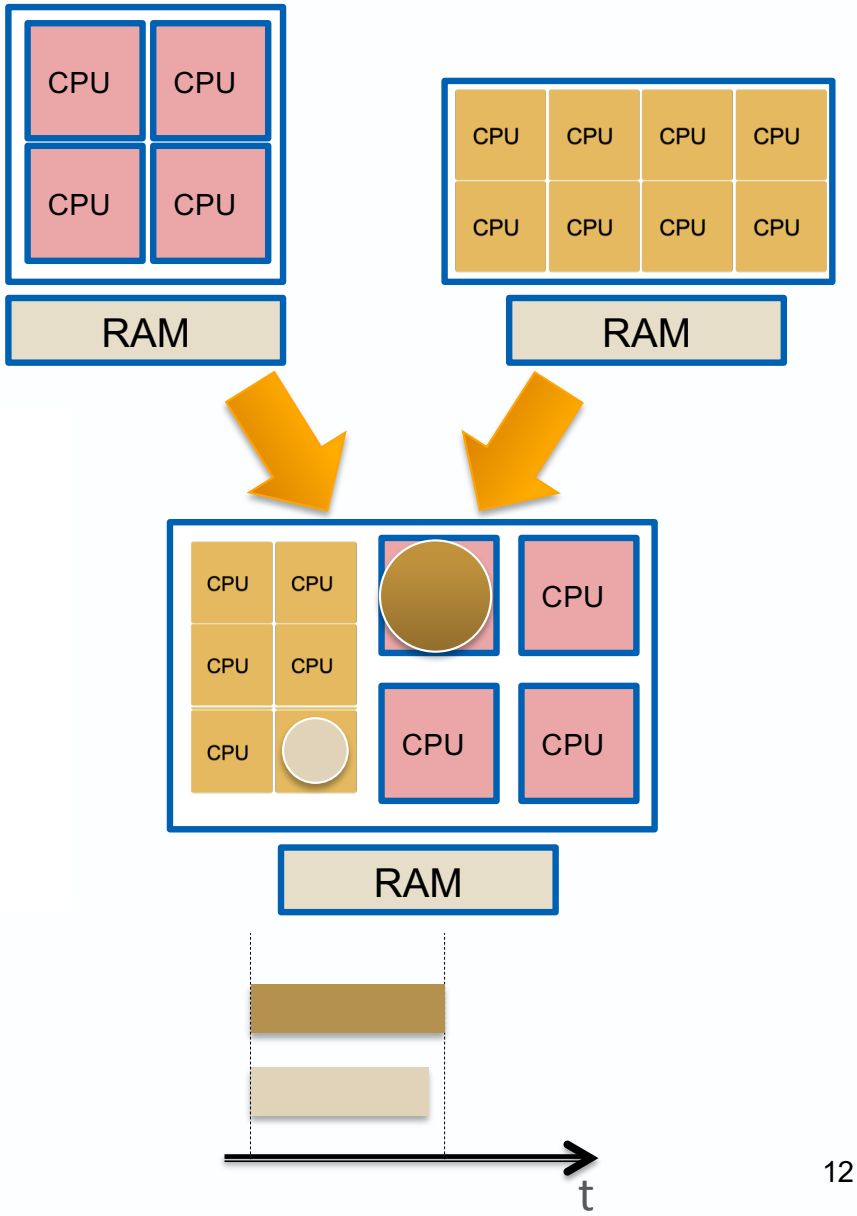


Heterogeneous multicores

Going HMP: *Hetergeneous Multi-Processing*

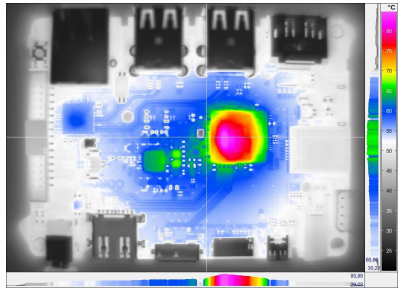
- Single ISA!
- Linux sees all cores
- Allows balancing workloads

Promising, still challenging



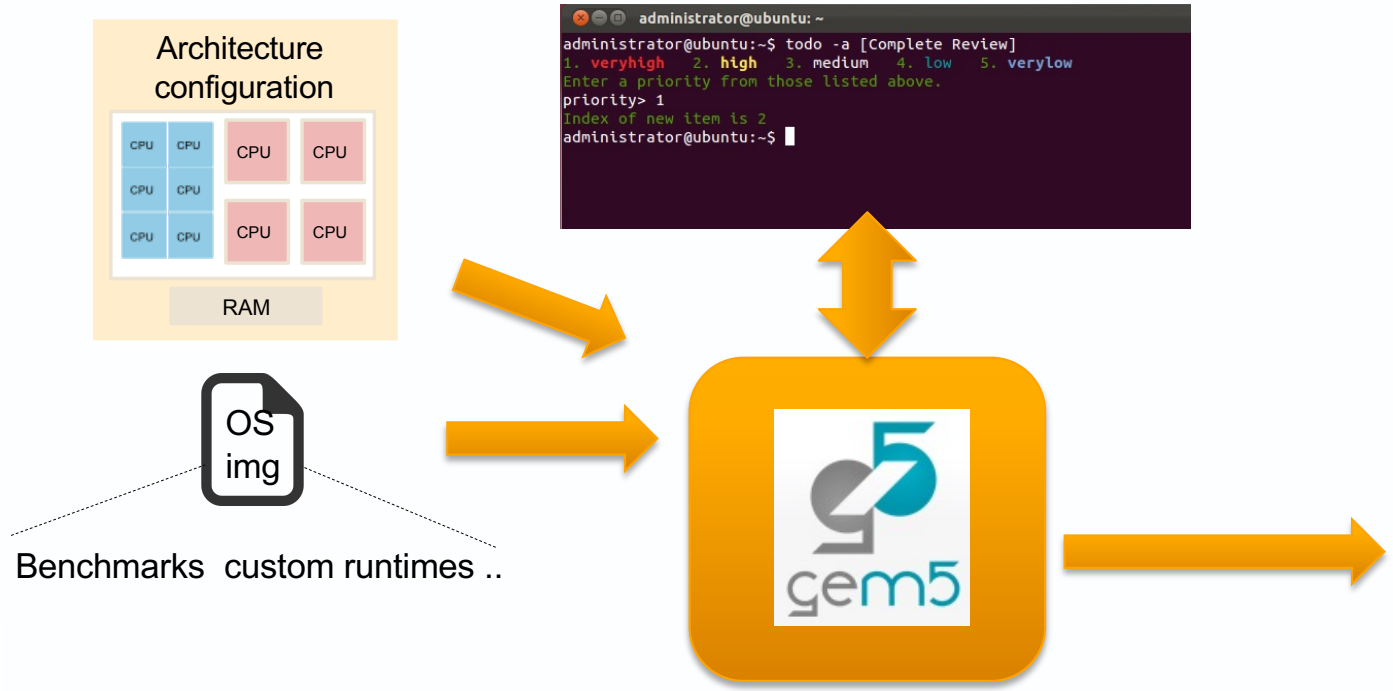
Measure & model architecture specifics

- Run benchmarks, measure performance & power
- Figure out architecture specifics & build power models

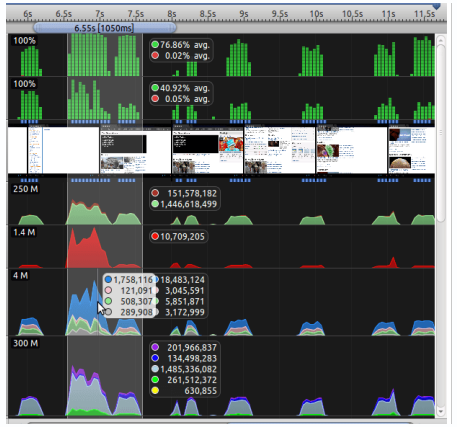


and then simulate various ideas

- Using Full-System simulators, i.e. gem5

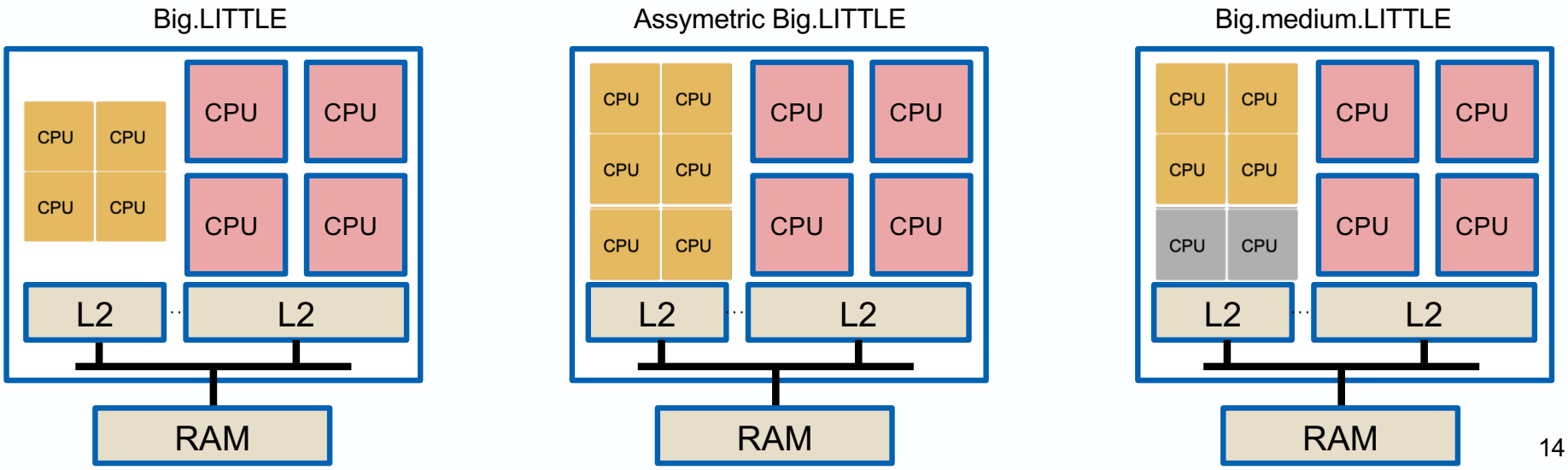


Exec statistics,
low-level behaviors,
POWER



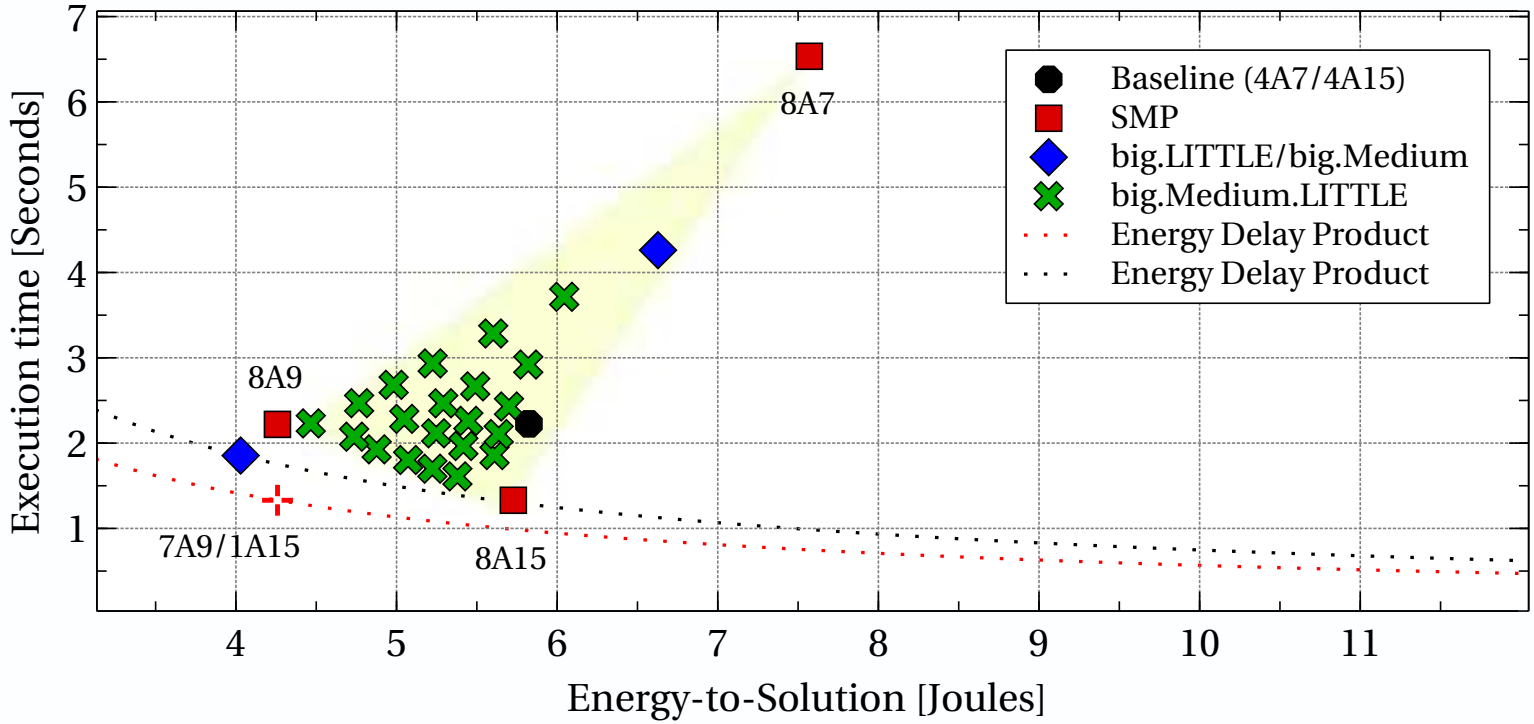
Playing with dozens of parameters for diverse blends

- Heterogeneity levels (big.LITTLE, big.medium.LITTLE..)
- Assymetry (core count per cluster)
- Cache sizes, coherence protocols
- Memory hierarchy: cache levels etc.
- Interconnect type (bus, mesh, torus…)
- Memory technologies
- …



Analyzing performance/power tradeoffs

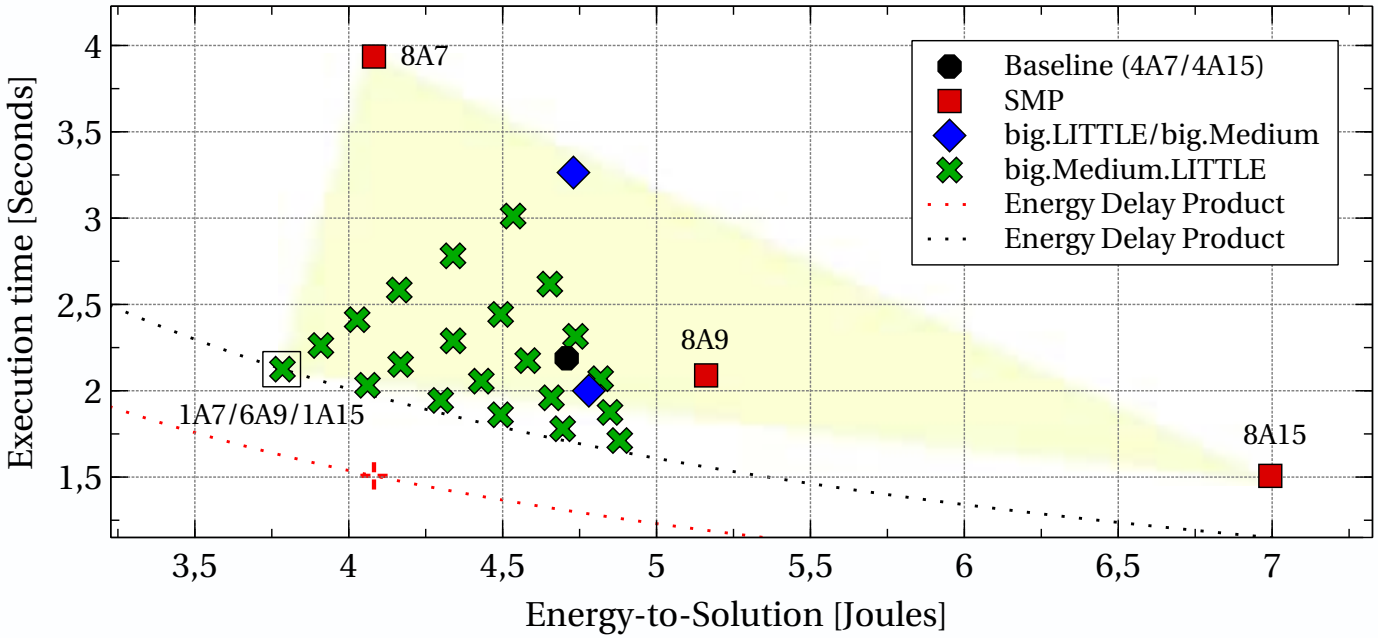
- 8 cores systems, either
  - Symmetric ■
  - big.LITTLE ◆
  - big.medium.LITTLE ✖



(f) *lud*

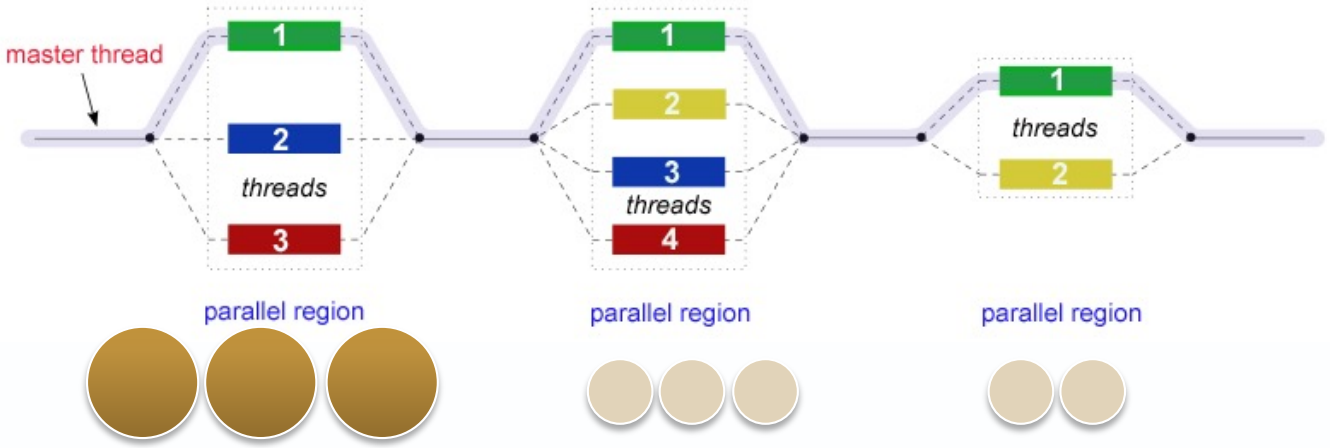
Analyzing performance/power tradeoffs

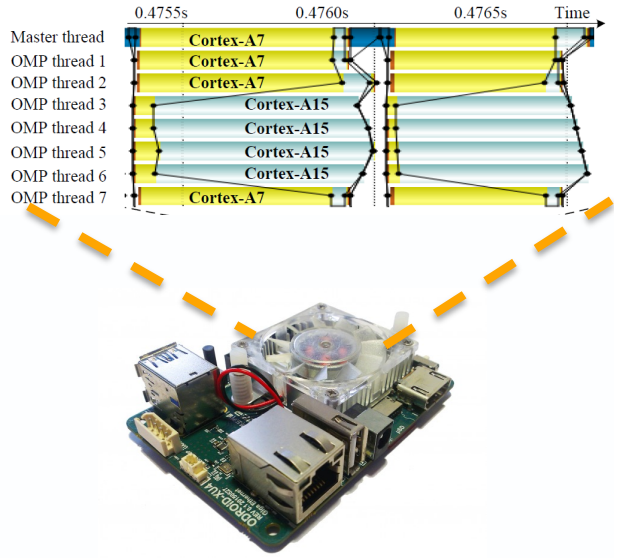
- 8 cores systems, either
  - Symmetric ■
  - big.LITTLE ◆
  - big.medium.LITTLE ✖



(e) *kmeans*

Most legacy codes use OpenMP < 4.0



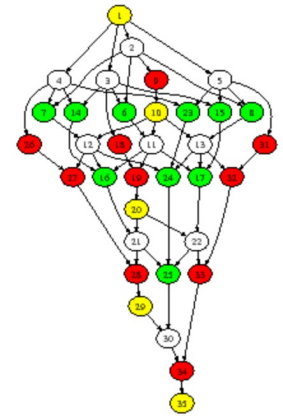


OpenMP 4.0 & friends (OmpSs) go tasking

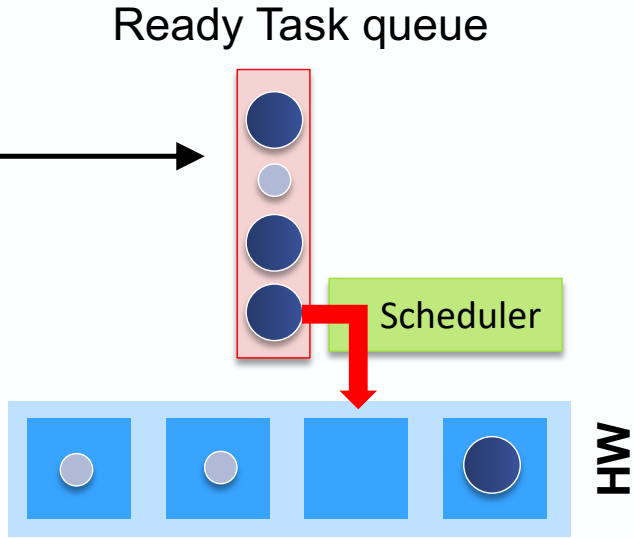
- OpenMP 4.0, OmpSs, pragma-based

```
#pragma omp task depend(in: x) depend( out: y) depend(inout: z)
```

- More parallelism available @run-time, opportunities for smart job scheduling

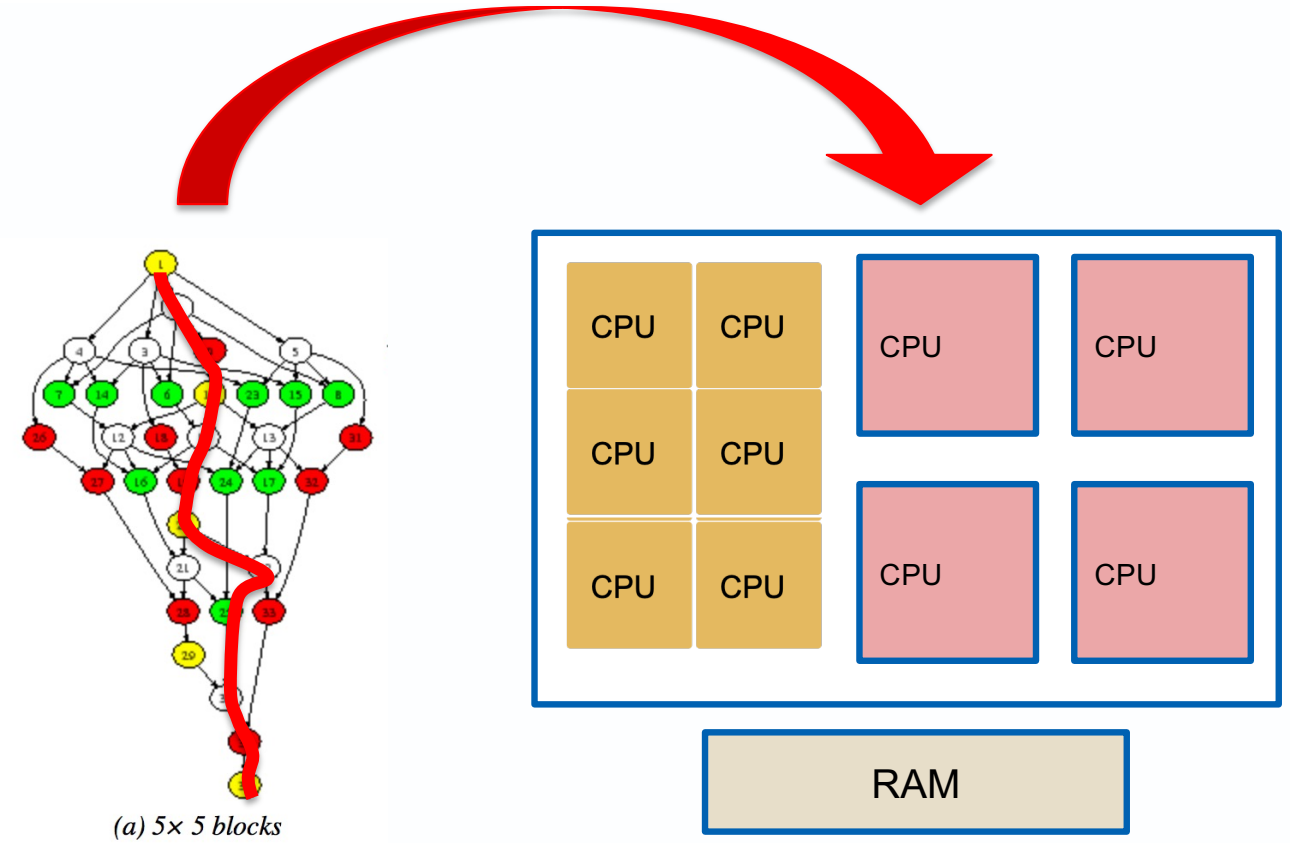


(a) 5× 5 blocks



Many opportunities

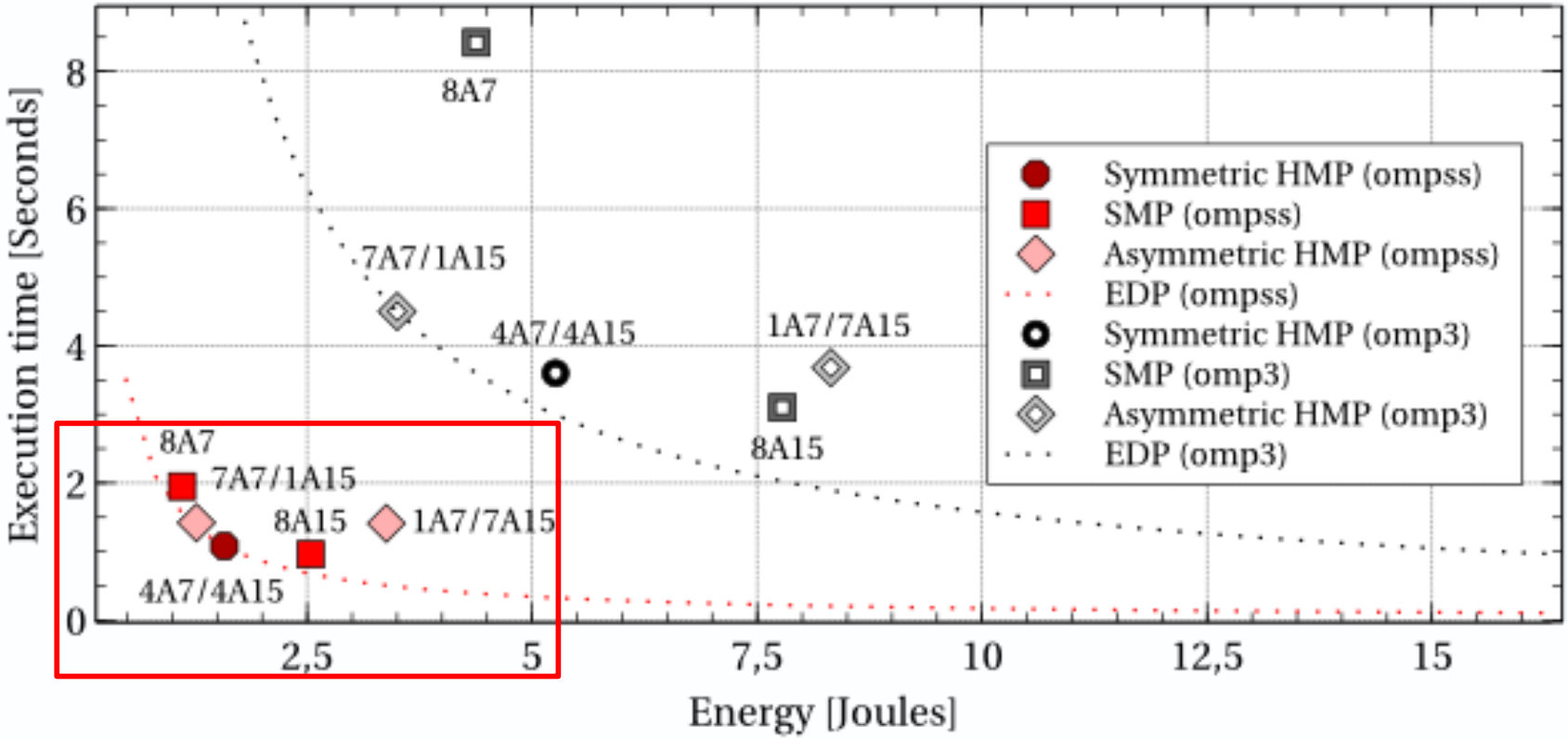
- Defining task queues



(a) 5× 5 blocks

OmpSs significantly outperforms OpenMP

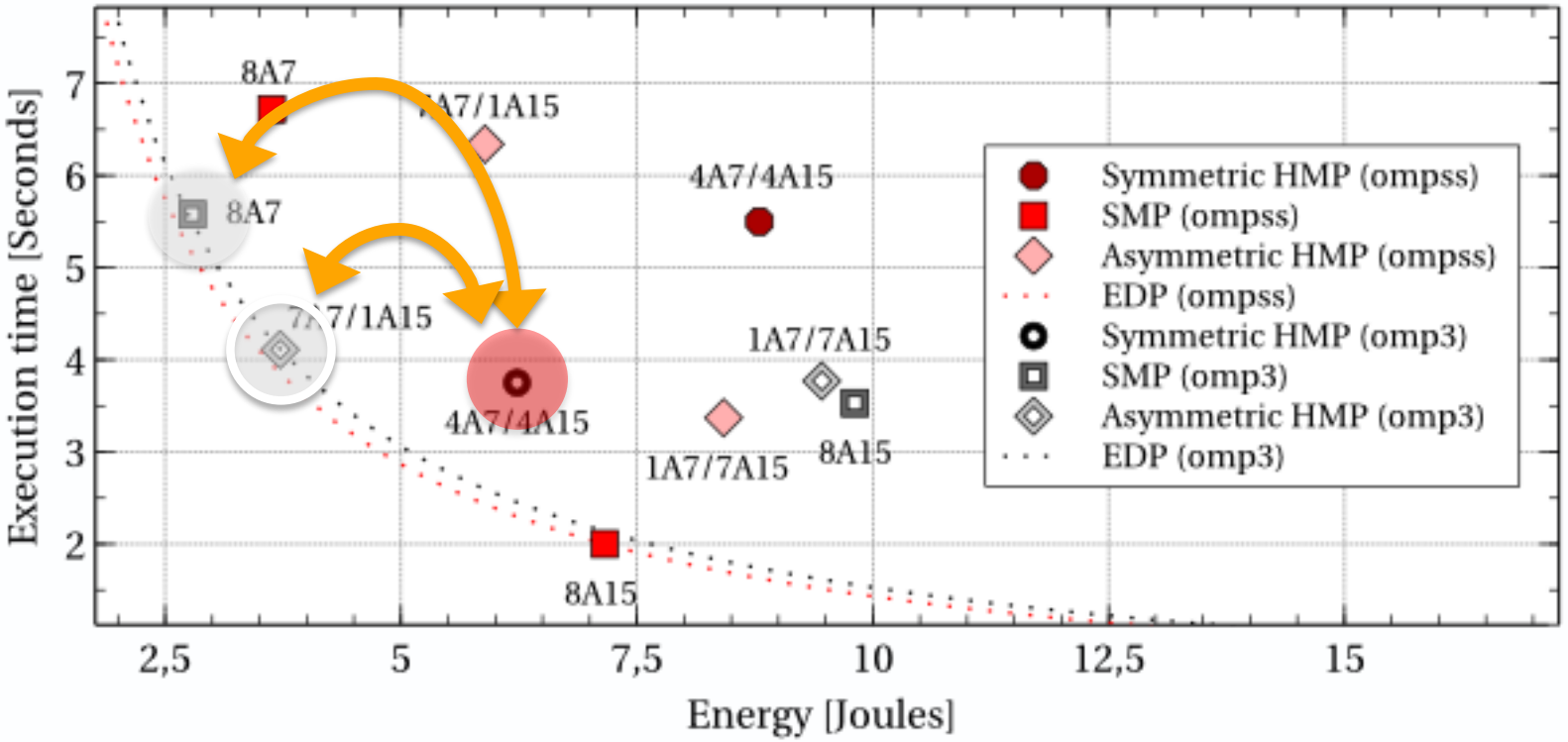
- LUD exhibits significant inter-thread sharing (row/column dependencies)
- OmpSs implicit data dependency management helps
- Most configurations line up on the same EDP, with tradeoffs



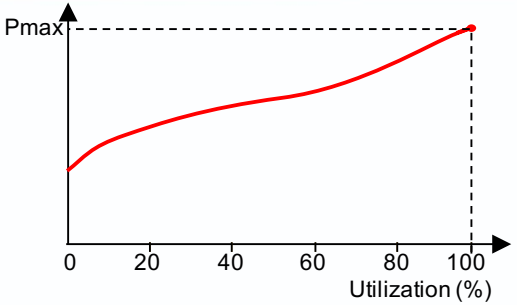
Not across all benchmarks!

Again asymmetric HMP configurations provide tradeoffs

- 4 A7/4 A15 has worse EDP compared to 8 A7
- 7 A7/1 A15 ~ 40% faster, ~40% more energy

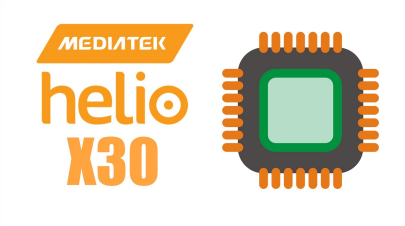


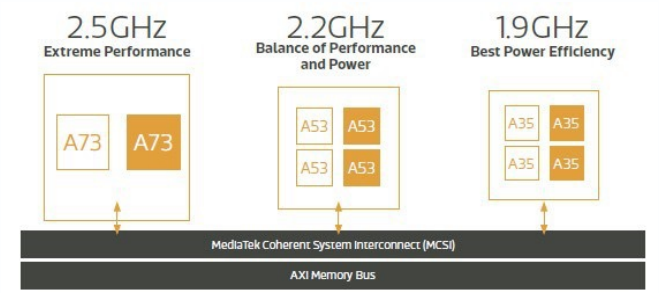
On static power

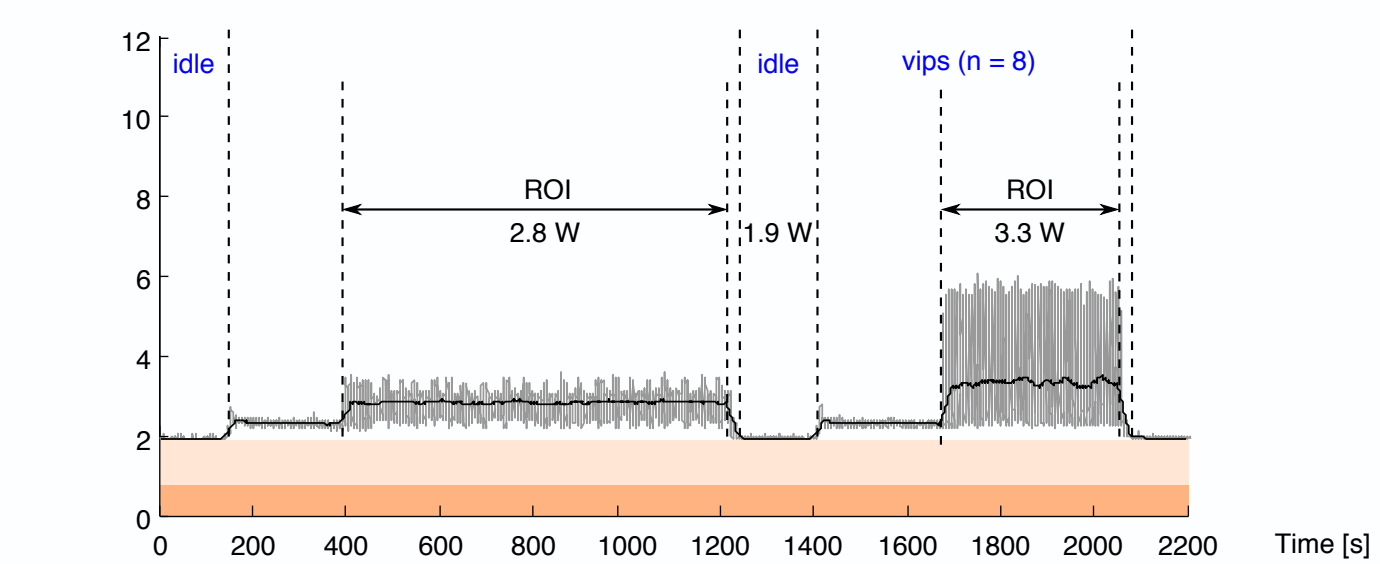


Still, chips remain vastly non energy-proportional

- Dynamic power: toggling not purely proportional to load
- Static power: « constant » price to be paid
  - For each transistor (each path from Vcc to GND)



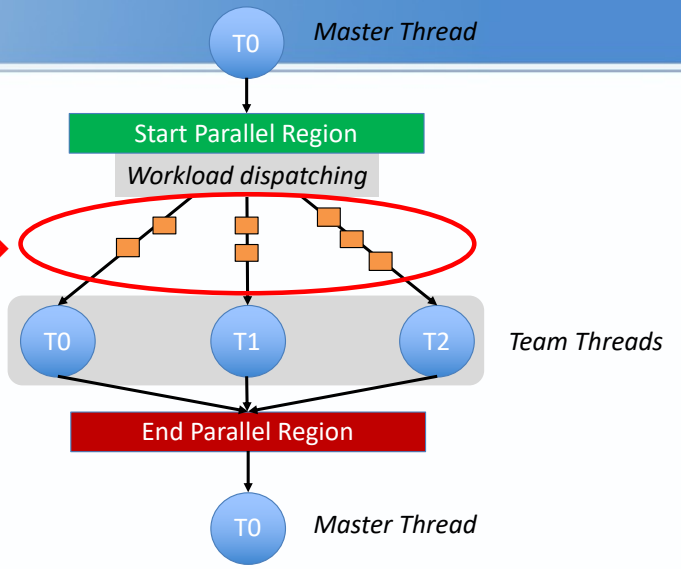




Acting upon, the levers

Going from ASAP to As (*Power Efficient*) AP

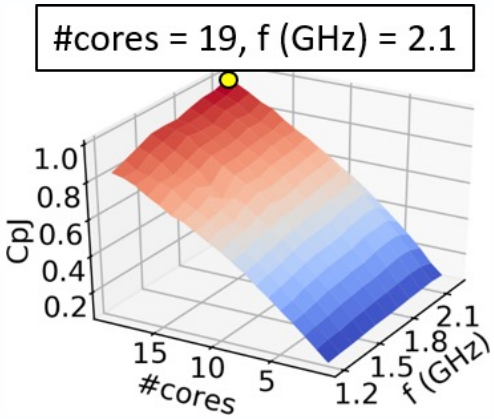
- OpenMP application progress can be tracked … chunks →
- Power available @ software-level
- Tweaked GOMP (GNU OpenMP) s.t. *runtime knows*



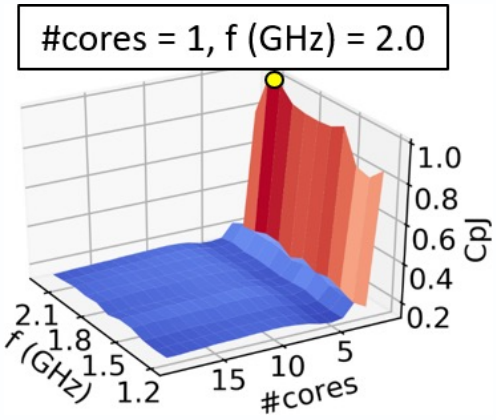
Devising a synthetic benchmark with knobs:

- CPU intensive / Memory intensive
- Influence of used *#cores* and *frequency*
- Measure energy efficiency (CpJ: Chunks per Joules)

CPU-Intensive



Memory-Intensive



Intel Xeon server = {#core, frequency} #core <= 19 1.2GHz < Freq < 2.1GHz

Acting upon, the levers con'd

Different mix

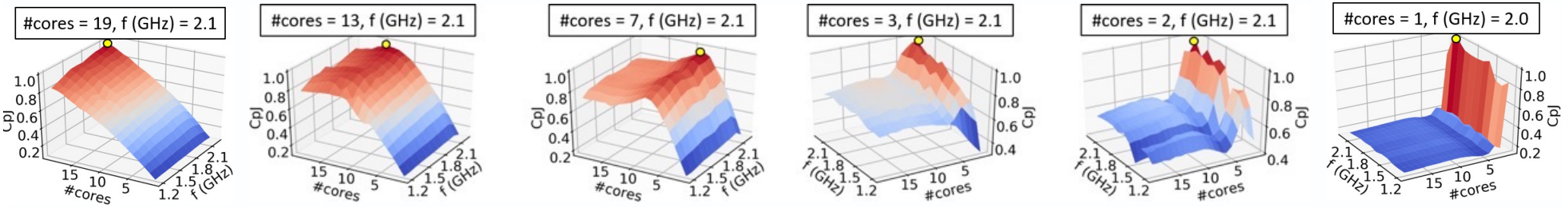
- Compared to standard Linux governors

Compute intensity Memory intensity

Benchmark		C100M0	C98M2	C96M4	C90M10	C80M20	C0M100
Best Conf.	CpJ	3866	2505	1581	818	517	336
i.e. Reference	CpS	303k	170k	90k	40k	25k	12k
vs.	CpJ	10%	25%	29%	95%	60%	442%
Performance	CpS	-12%	-11%	-19%	21%	1%	151%
vs.	CpJ	16%	24%	33%	44%	75%	469%
Powersave	CpS	75%	59%	49%	56%	92%	355%
vs.	CpJ	10%	20%	32%	18%	75%	433%
Ondemand	CpS	-12%	-14%	-17%	-1%	50%	193%
vs.	CpJ	10%	20%	29%	32%	56%	469%
Conservative	CpS	-12%	-14%	-19%	-16%	1%	160%

CPU-Intensive

Memory-Intensive

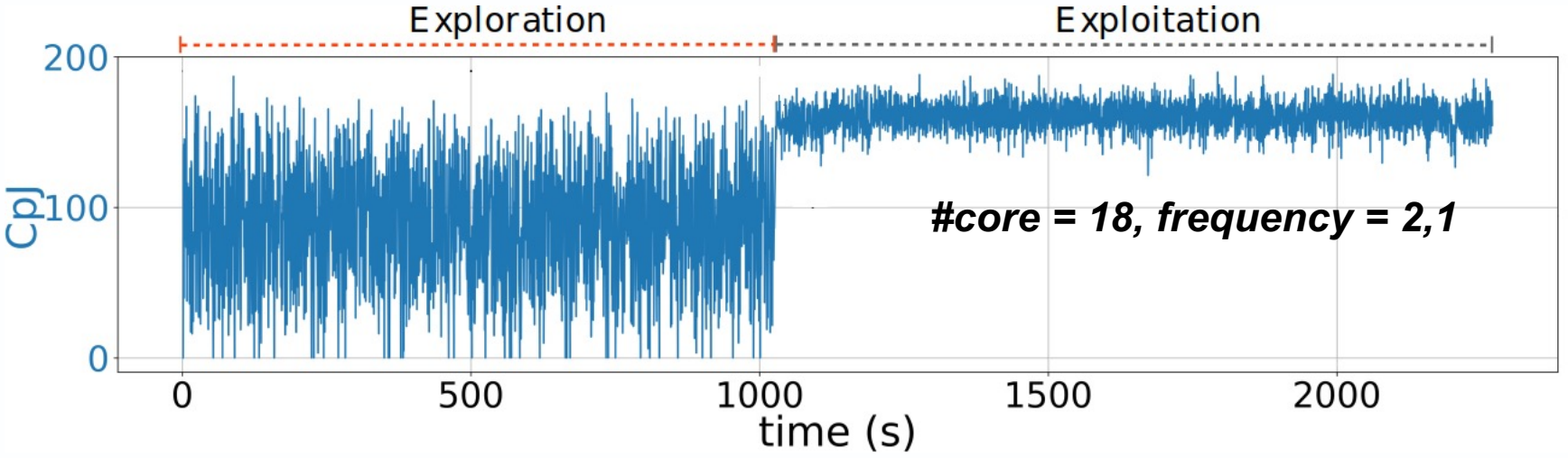


Intel Xeon server = {#core, frequency} #core <= 19 1.2GHz < Freq < 2.1GHz

Acting upon, the levers con'd

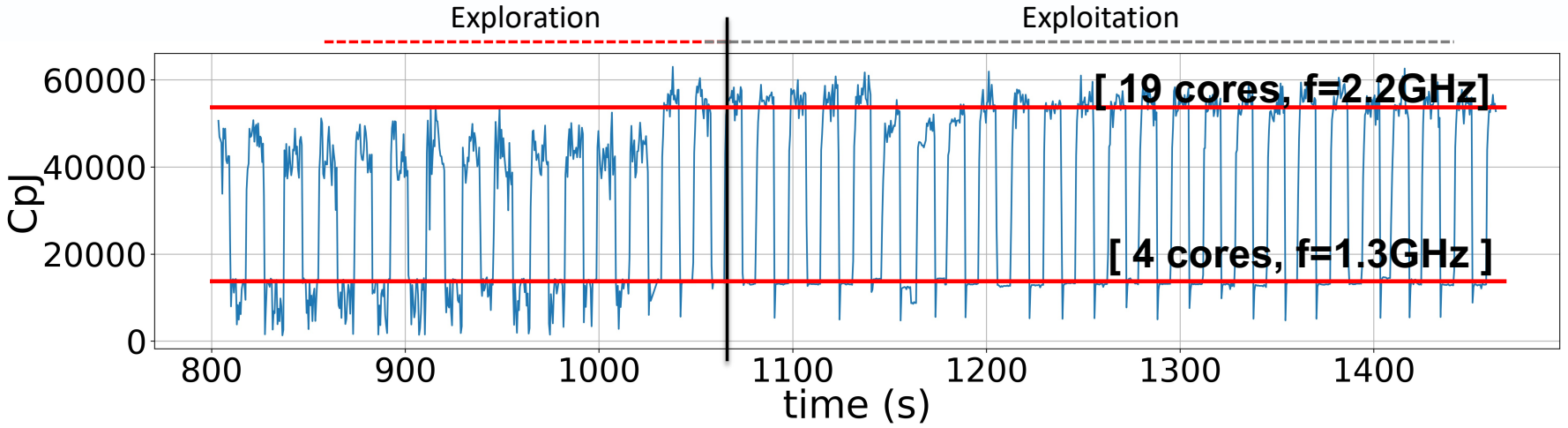
Getting the system to decide @ run-time

- As applications have phases, so assuming phases are
- Using reinforcement learning (overkill)



DGEMM

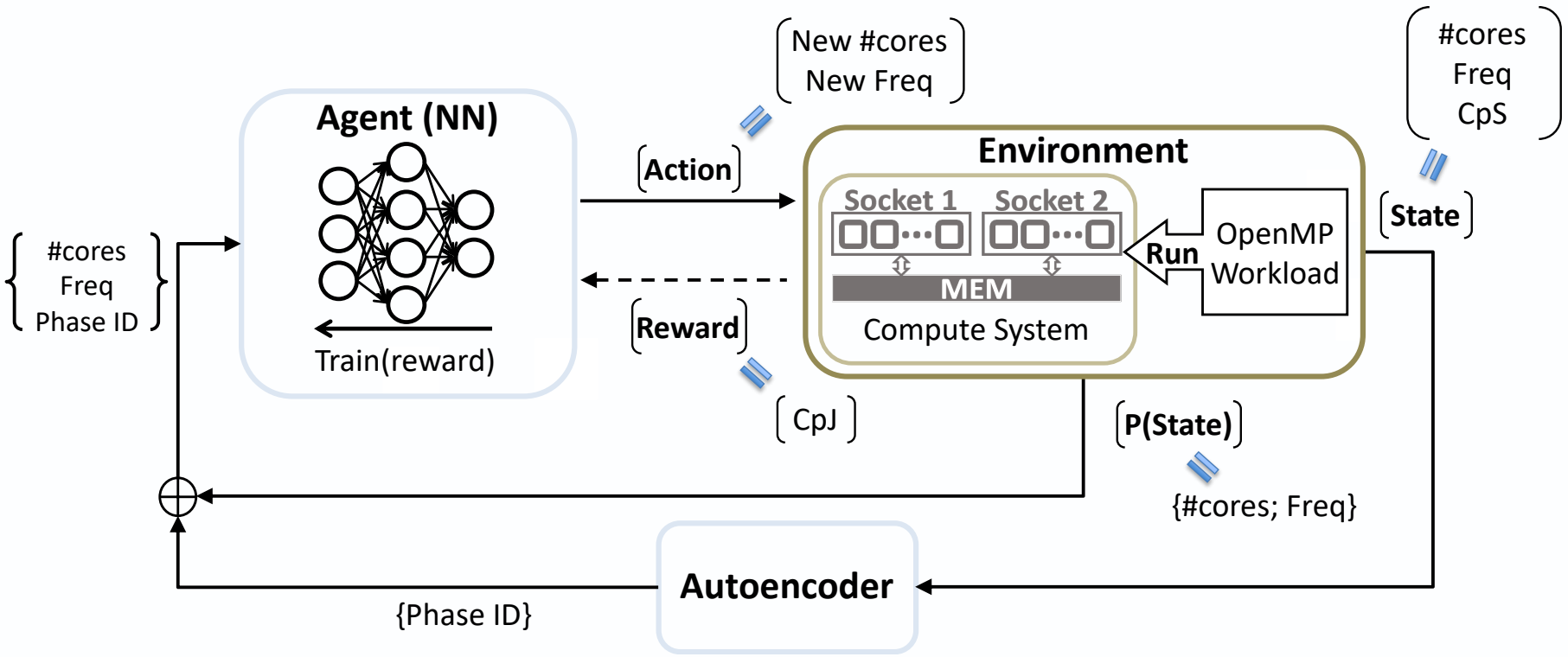
- With phases



Behind the scene

RL-based engine

- NN for reward prediction (DQN style)
- Tuned AE arch. trained for extracting phase information

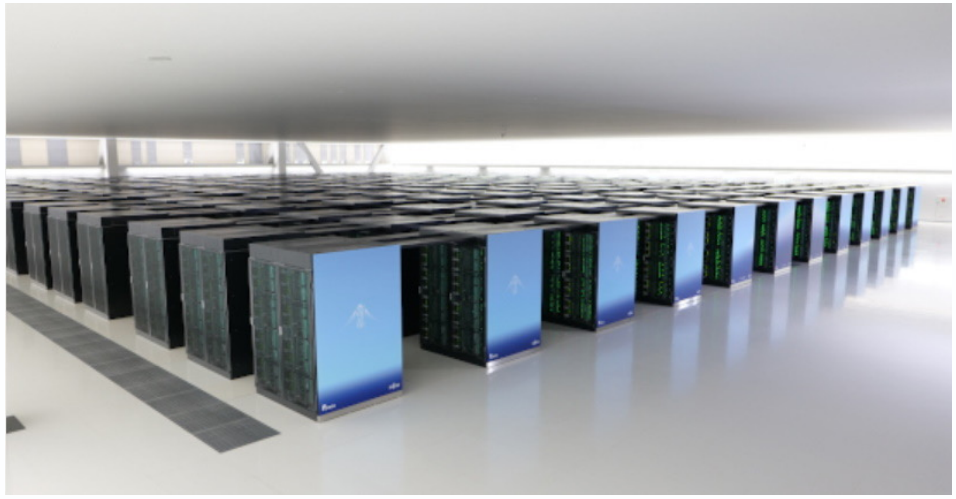


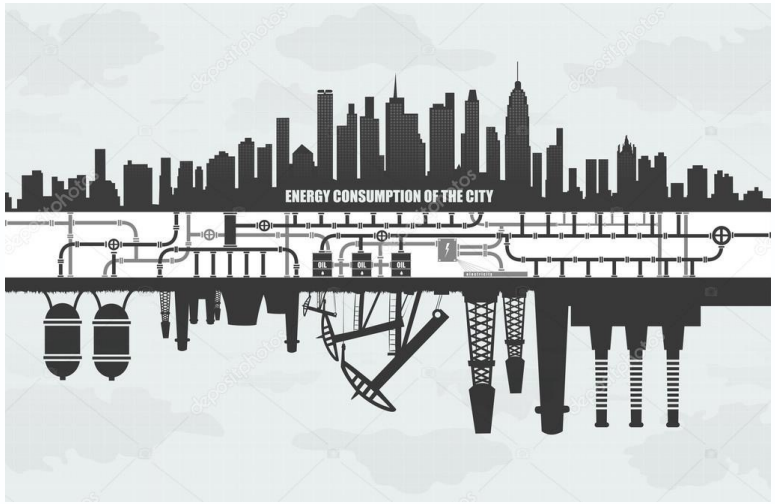
How are we doing power-wise?

Remember that (nasty) assumption?

Rank 1:

- Sunway TaihuLight: 442 PFLOPS (Riken, JP)
  - 159k nodes : ARMv8 64b @48+2 cores
  - 7,63.10^6 cores
  - **~30 MW – 3500€/hour operation**





Exploring a theoritical exaFLOP supercomputer

10^{18} FLOPS within **20 MW** power budget **-> 50GFLOPS/W**

- ≈ 30-40% for CPUs, rest on interconnect, storage & losses in power distribution

Assuming same CPUs

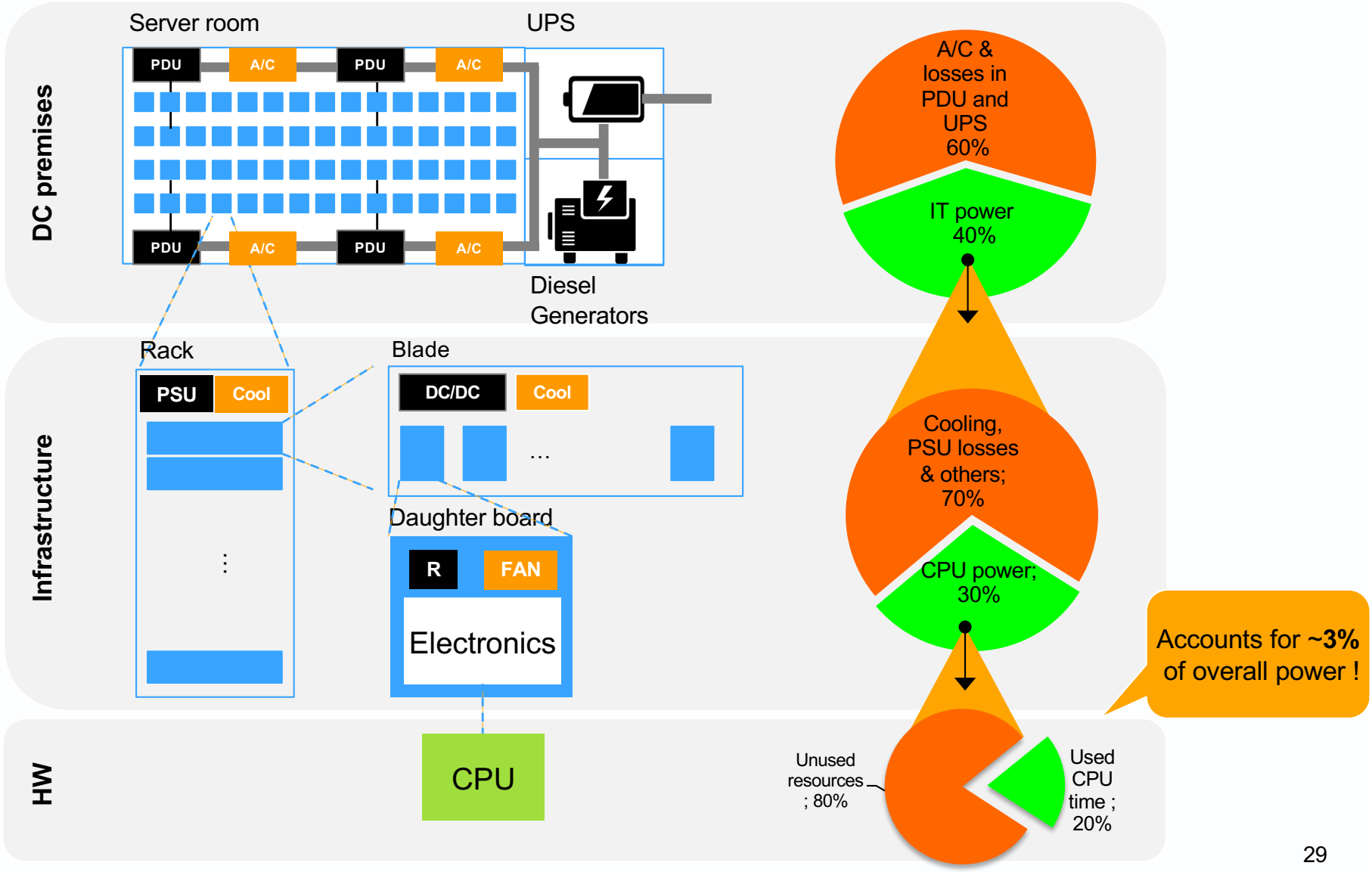
- Twice the performance, twice the CPUs (optimistic) → 15M CPUs

That makes 6MW for 15M CPUs

- 0.4W / CPU …

Where do kWh go?

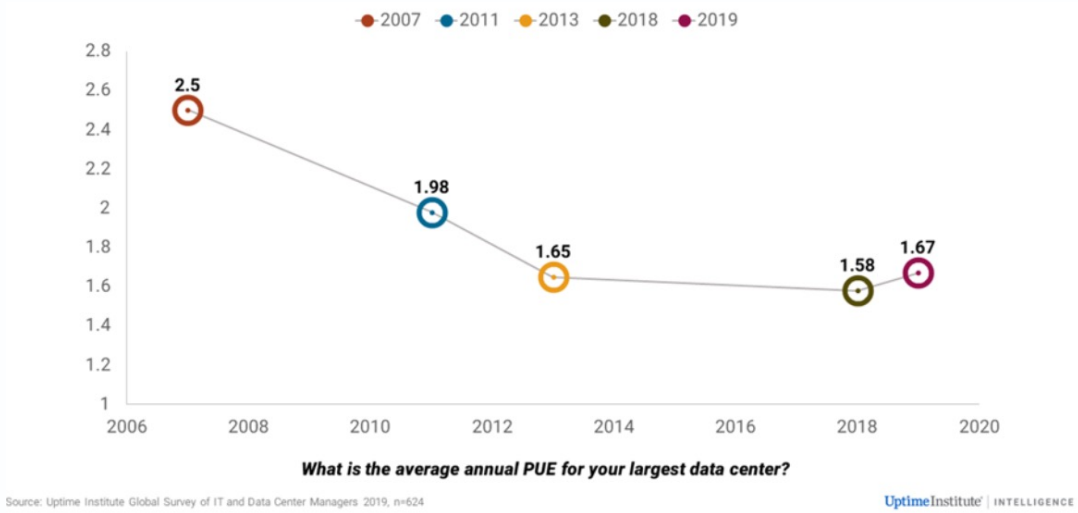
Improving energy efficiency @ 100% load?



Everything PUE

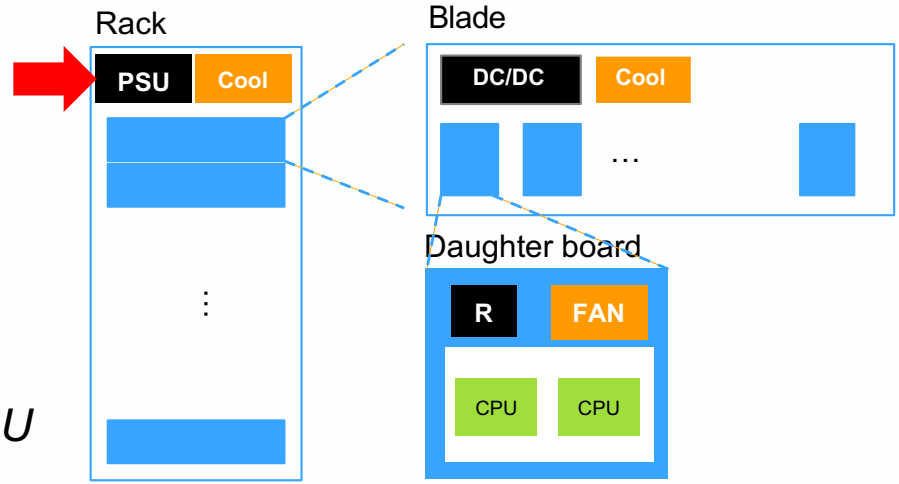
Power Usage Effectiveness (PUE ISO)

$$PUE = \frac{\text{total facility power}}{\text{IT equipment power}}$$



- Behind the scene it's not what you may *think*

$$PUE \neq \frac{P(\text{🏢})}{\sum P_i(\text{CPU})}$$



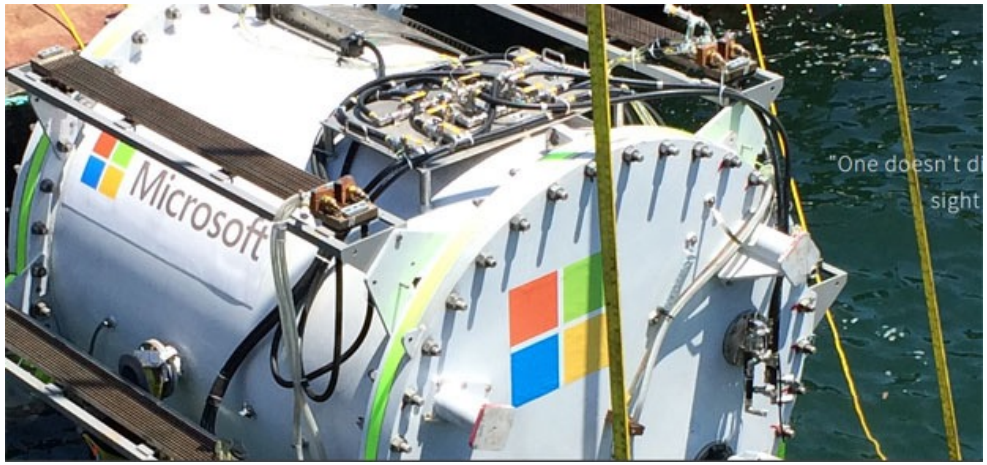
- *There's PUE cat. 1, cat. 2, cat. 3…*
  - *Cat 3: IT power measured @ in-rack PDU*

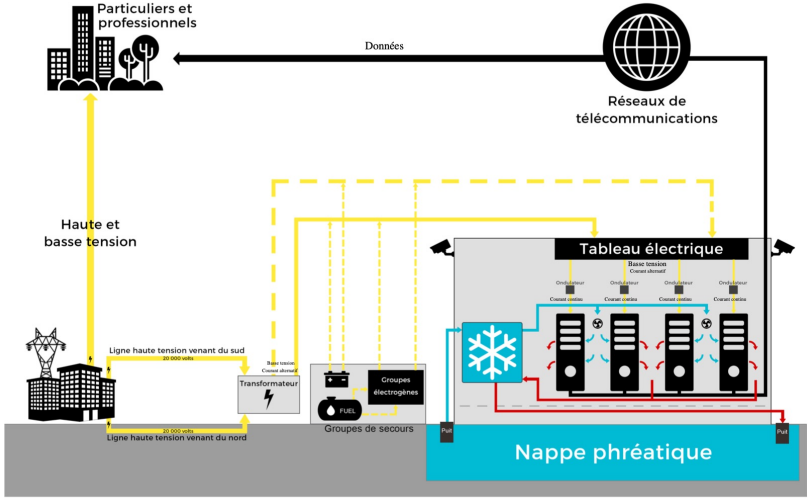
https://www.scaleway.com/en/transparency/

Everything PUE

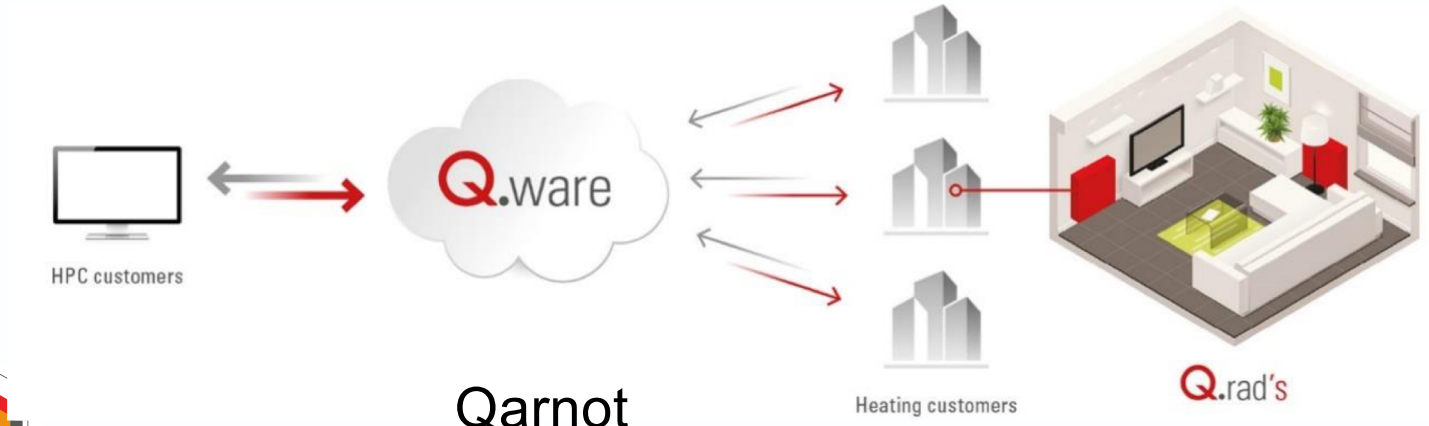
In the news

- Underwater (Microsoft), tapping into ground water, growing CO_2 absorbent algae





- Waste heat reuse
  - Compute with a temperature setpoint



Qarnot

Defab

And renewables?

Datacenters have unparalleled energy density

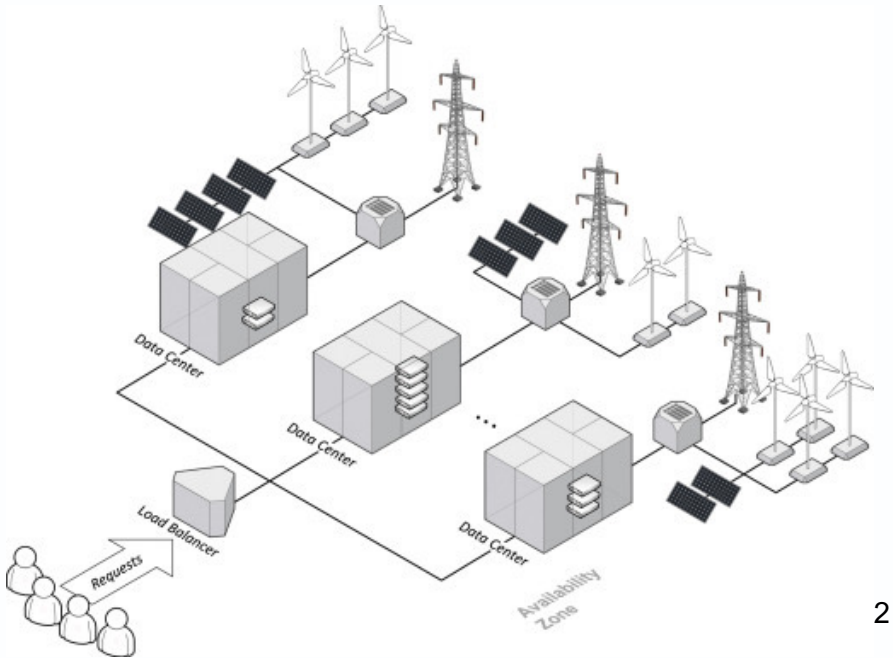
- 1 rack ~ 1m^2 → over 10kW
- Scaleway DC5: 10.000m^2, 20.5MW → 2kW/m^2
- Solar Peak ~ 250W/m^2

« Virtualized » carbon-neutral Data Centre via **compensation**

- GAFAM into « Renewable Energy Credits », financial assets linked to renewable energies
- However announced intent to match DC power w. renewables (hour basis)

Trend is more into energy mix

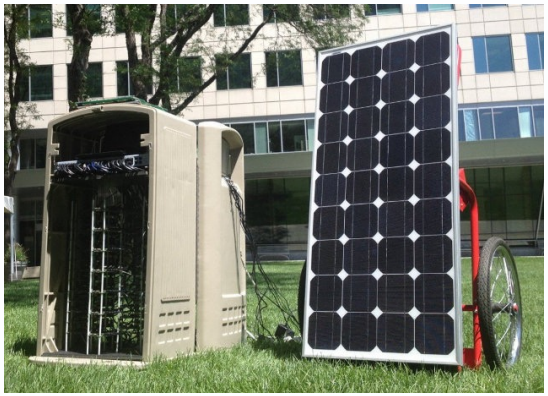
- Interesting optimization questions
- Moving workloads across the globe

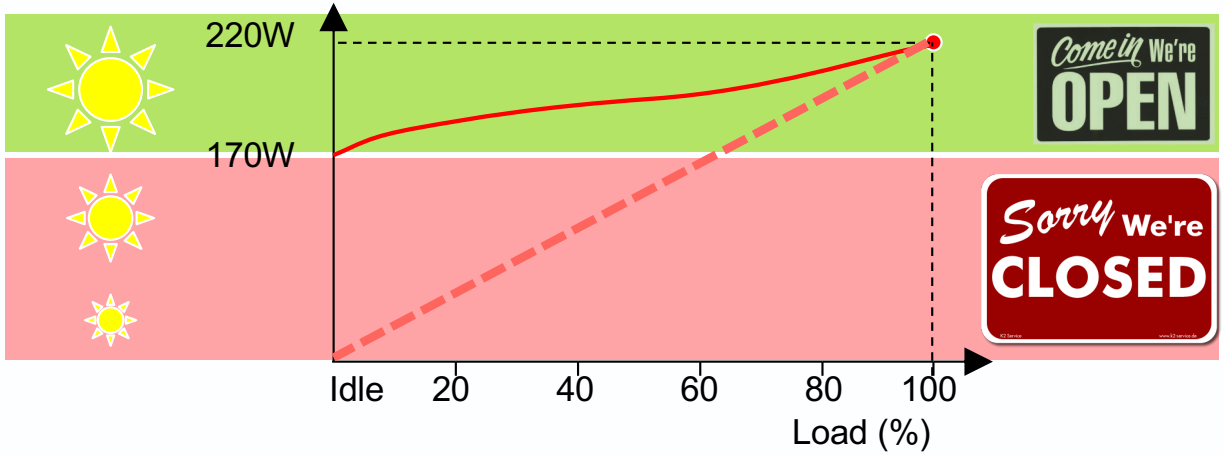


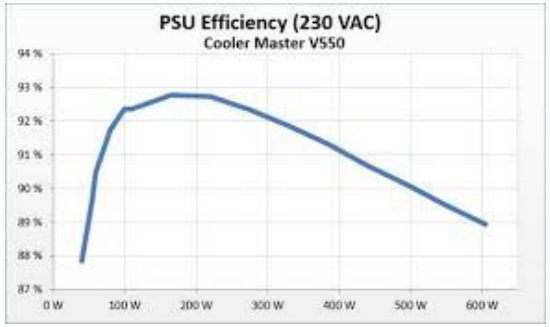
And renewables? Cont'd

Very little on compute equipment powered straight from the sun

- Google « solar cluster »
- 48-node ARM cluster (MIT)





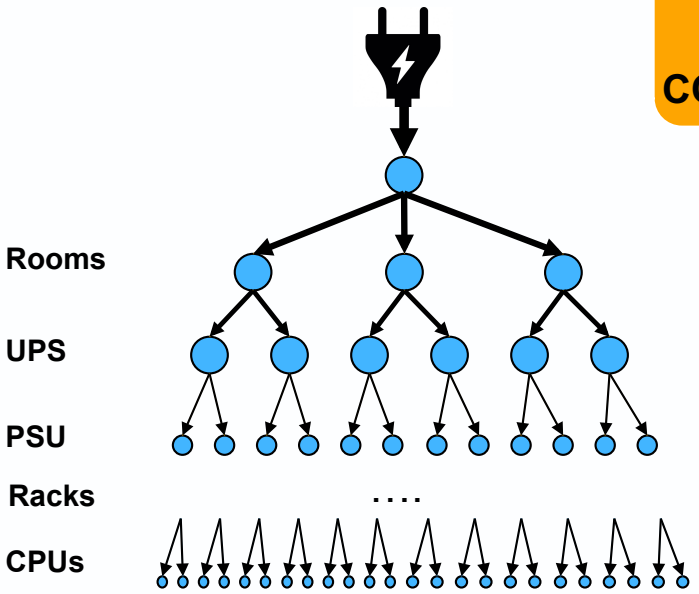


Upside down

Questionning the « producer / consumer » approach

- Energy harvesting is distributed in essence (solar)
- Centralizing and then re-distributing energy incurs many losses
  - Distribution (transport)
  - Conversions AC/DC
  - Heat density requires cooling

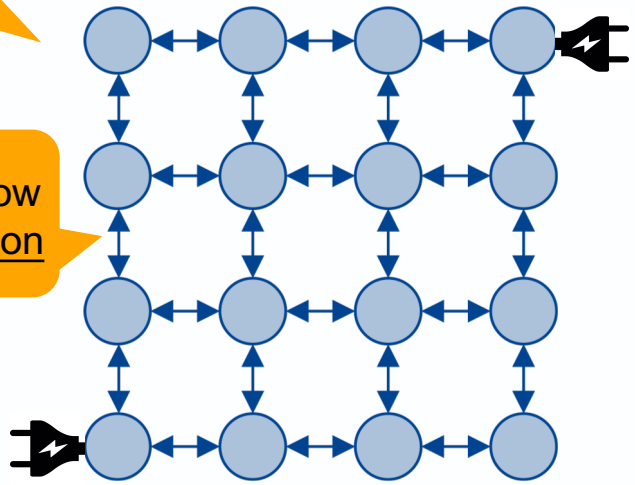
Typical power topology:
Fat-tree & one-way



What we're looking for:
Arbitrary & **TWO**-way

NOT a comm. network:
Both
COMM and **POWER** network

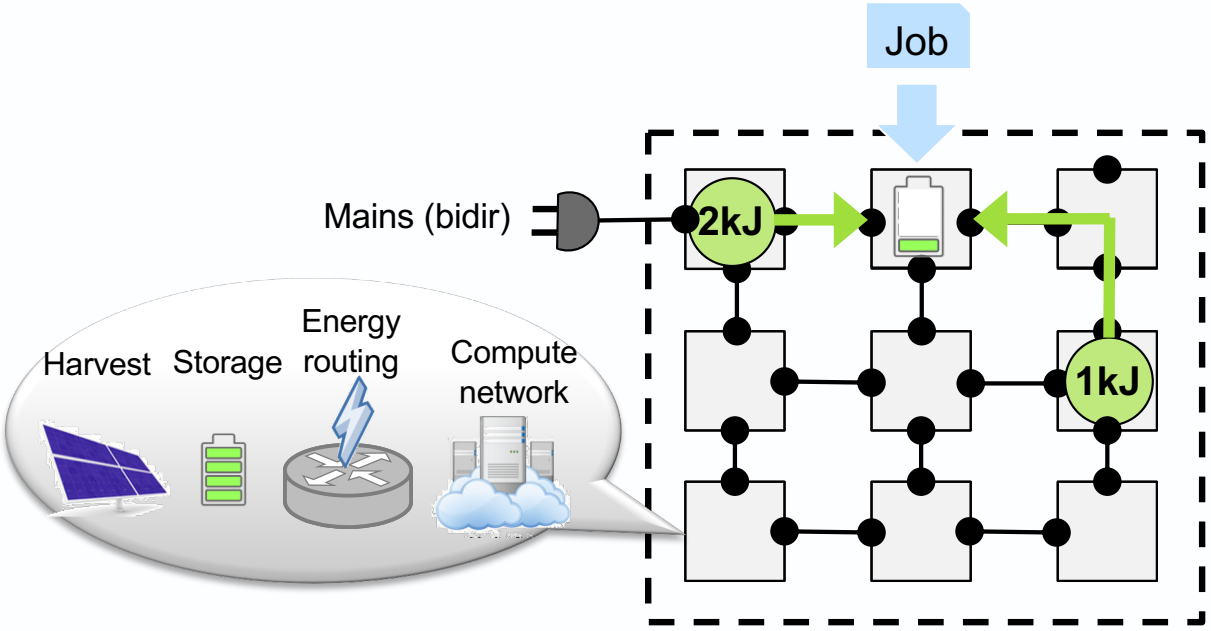
Both **DATA** and **POWER** flow
in a Software defined fashion



Key principles

Energy, compute & memory are similar allocable resources!

- Energy harvested, stored and transported where necessary in the cluster
- No AC/DC conversion, no UPS: no losses in PSUs, passive cooling
  - PUE ~ 1
- Extreme reliability due to software-defined power supply network
- Excess energy pushed back into the grid



What to expect

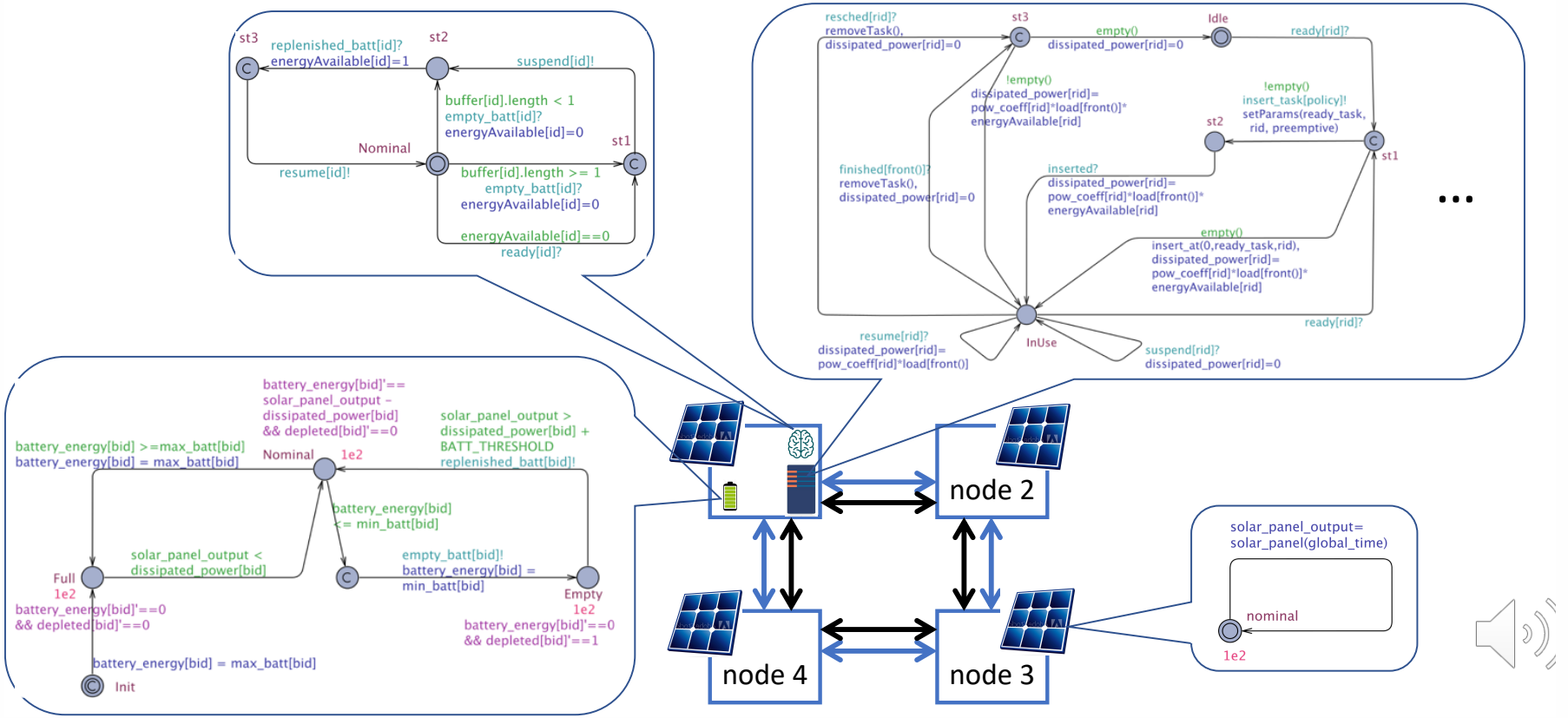
Distributed energy-driven system (no grid)

- Modeled as Stochastic hybrid (timed) automata
- On a real-time benchmark (11 tasks)

Edges with discrete probabilities

Clocks as ODEs in different states, e.g.

nrj' == solar_power – dissipated_power

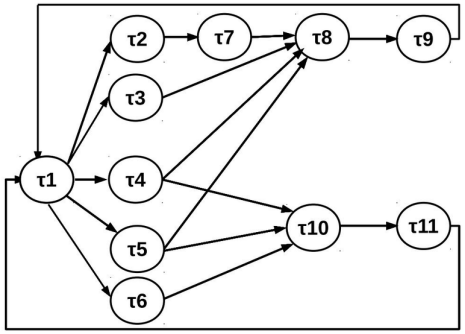


Abdoulaye Gamatié, Gilles Sassatelli and Marius Mikučionis. '**Modeling and Analysis for Energy-Driven Computing using Statistical Model-Checking.**', *Design, Automation and Test in Europe Conference - (DATE'21)*, virtual event, February 2021.

What to expect

Distributed energy-driven system (no grid)

- Modeled as Stochastic hybrid (timed) automata
- On a real-time benchmark (11 tasks)



Pr[<= 2*DAY] (<> exists(i:t_id) Task(i).Error)

Min. battery size **without** energy transfer

	B0	B1	B2	B3	B4	Total	Total homogen.
June	4.7	16.8	1	1	1.7	25.2	84
December	6.7	23.4	1.7	1.7	3.3	36.8	117

60%

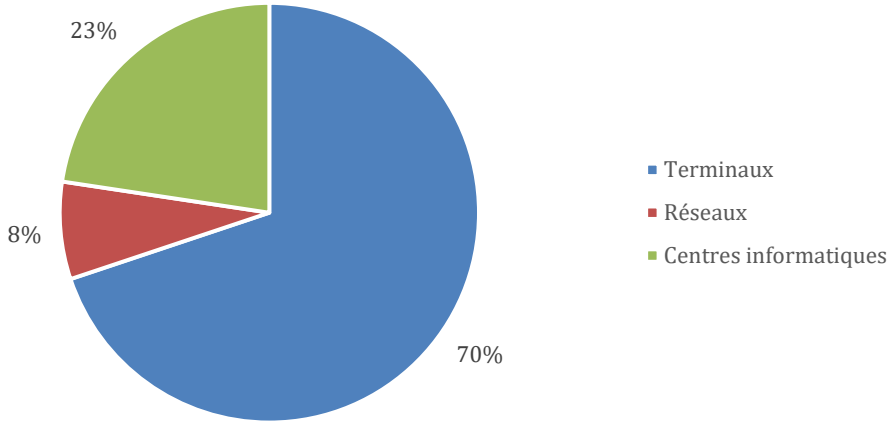
Min. battery size **with** energy transfer

# Nodes	B0	B1	B2	B3	B4	B5	B6	B7	Total.
6 (het.)	5.8	10	5.8	5.8	6.8	4.2	–	–	38.4
8 (het.)	5	6.8	4.2	4.2	5.5	4.2	4.2	4.3	38.4
8 (hom.)	5	5	5	5	5	5	5	5	40

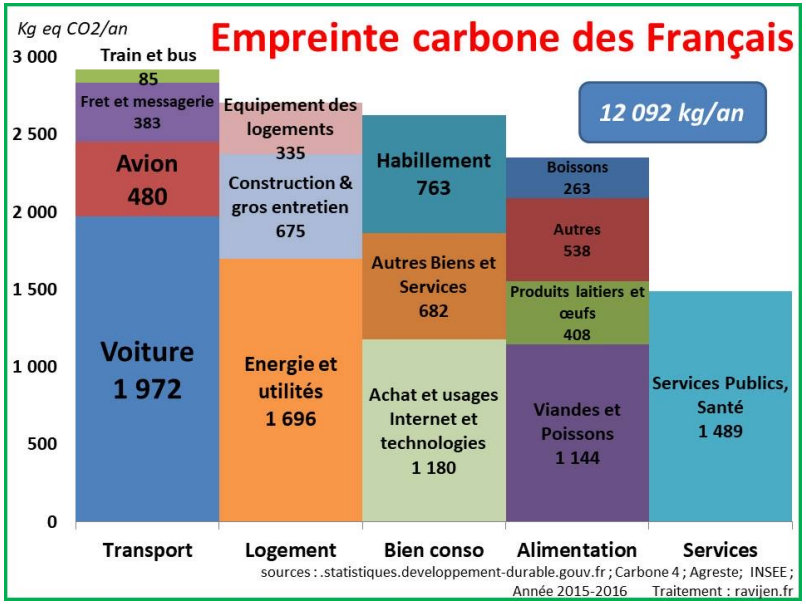
Summing up

With a grain of salt.. *It's all estimates & not CO_2 abstractable*

- DC global energy consumption said[1] to have rosen 6% only since 2010 (550% compute)
- Which is itself only a fraction of the IT power consumption
- …itself not much compared to our CO_2 footprint



« ÉTUDE RELATIVE À L'ÉVALUATION DES POLITIQUES PUBLIQUES MENÉES POUR RÉDUIRE L'EMPREINTE CARBONE DU NUMÉRIQUE »
http://www.senat.fr/rap/r19-555/r19-555-annexe.pdf



http://ravijen.fr

Plenty of good reads

- The Shift Project / J.M. Jancovici, V. Courboulay, Ademe…
- GDS EcoInfo, https://labos1point5.org, https://csi-ins2i.cnrs.fr …
- https://www.electricitymap.org/ …

1: E. Masanet et Al., « Recalibrating global data center energy-use estimates », Science, 367(6481), pp 984-986.

THANKS!