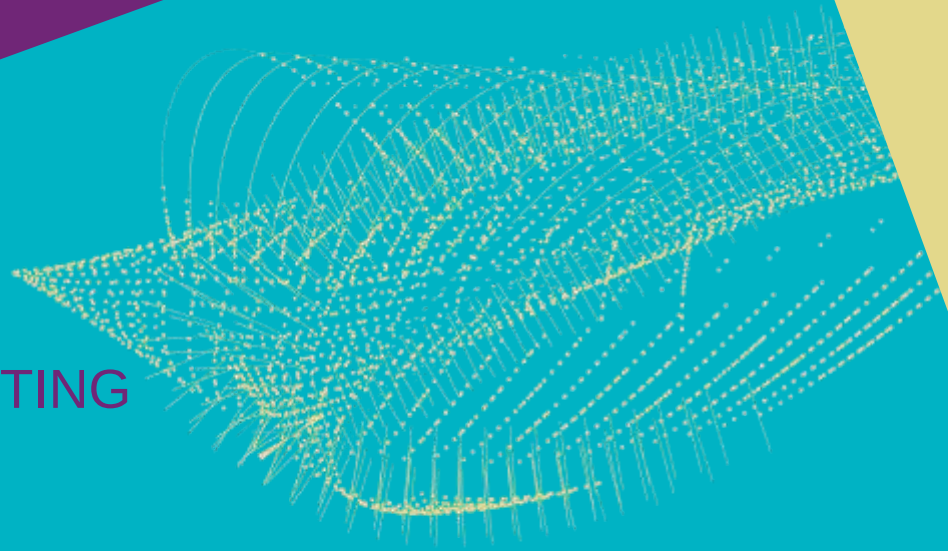
Nikhef

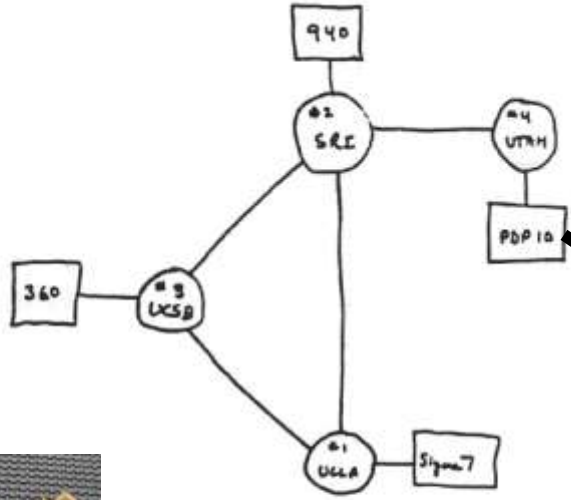


PDP – ADVANCED COMPUTING

SMALL IS BEAUTIFUL? TALES OF SCALING …

David Groep
Jamboree 2019

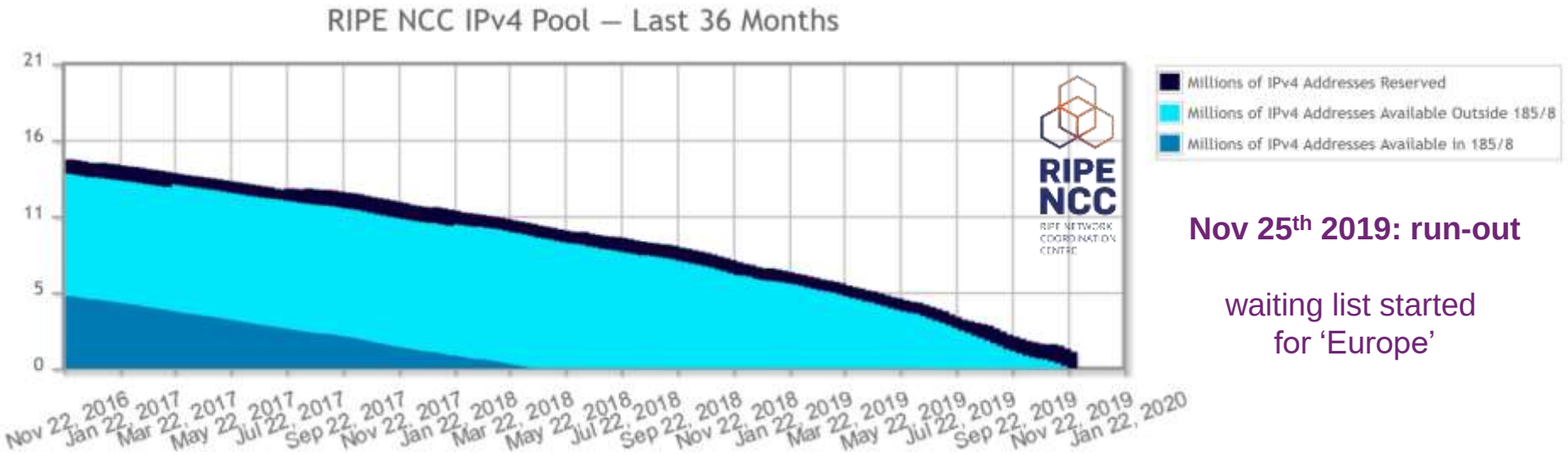
IT HAS BEEN 50 YEARS SINCE THIS WAS SMALL …



‘YE OLDE COMPATIBILITYE’

Image source: Alex McKenzie and "***Casting the Net***", page 56. See https://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/atlas/arpanet2.gif ; acoustocoupler: Wikimedia

SMALL NO MORE: ‘LEGACY IP’ (IPv4) HAS RUN OUT!



Nov 25th 2019: run-out

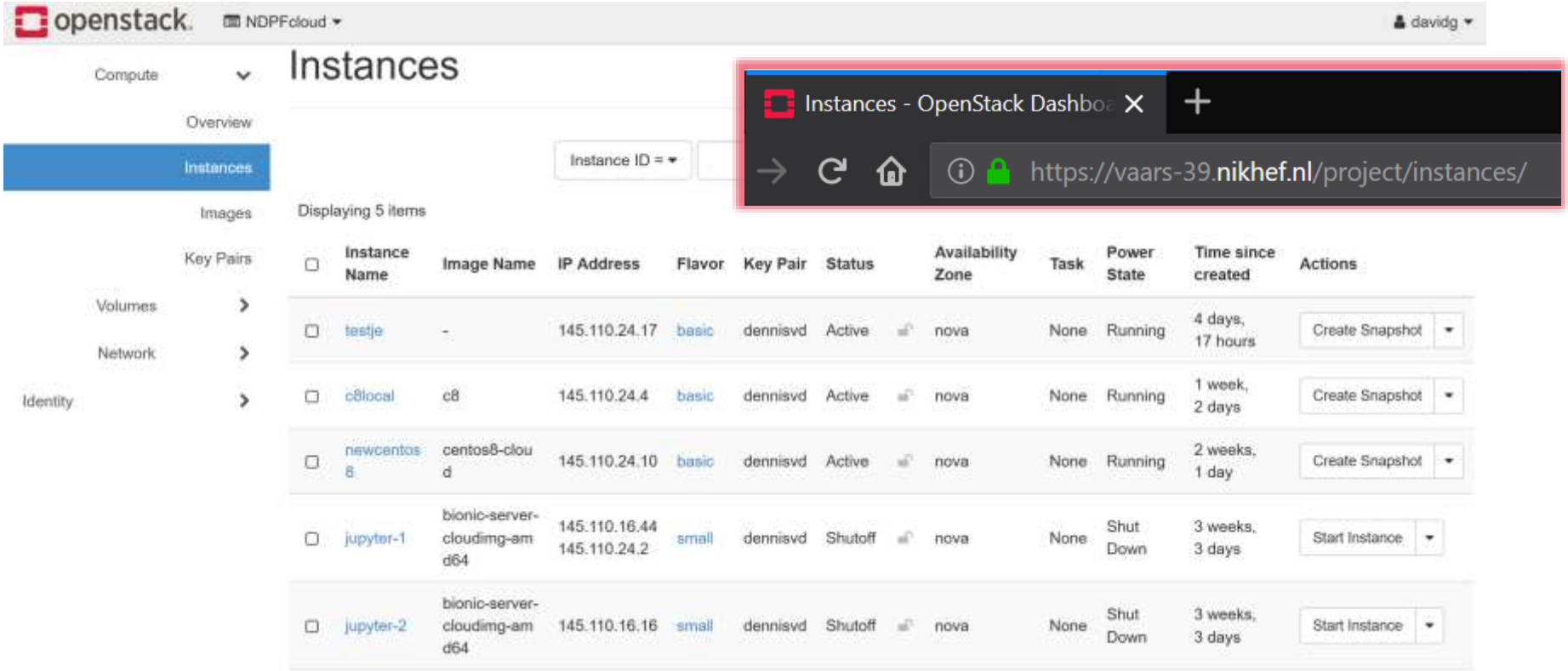
waiting list started
for ‘Europe’

					ffff	**192.16**	**185.**42
2a07	**8500 .. 8507**	0120	d100	e978	9eed	120c	89c1

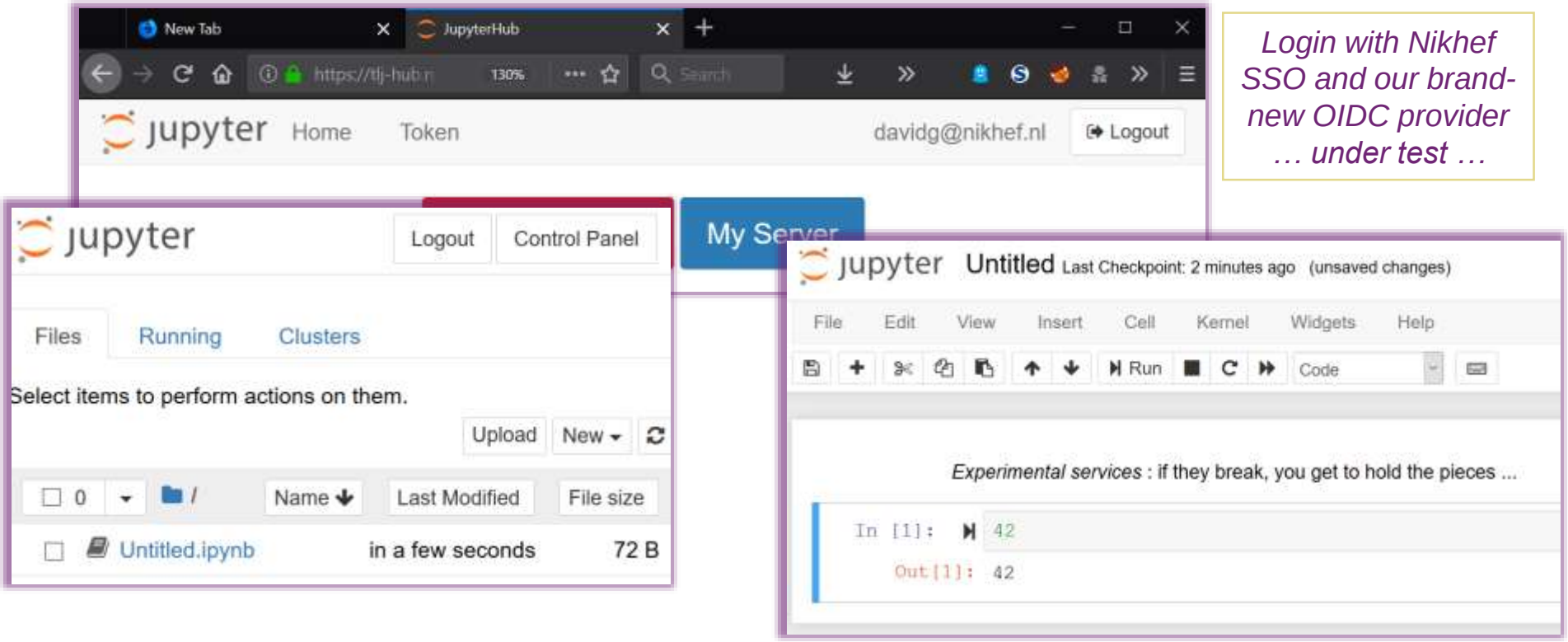
x ~ $8x10^{28}$ addresses
x ~ **$8x10^{16}$** networks

Nikhef

BUT WE STARTED EARLY AND CAN NOW DO: CLOUDS



FOR LOCAL USERS, FOR REMOTE USERS, FOR …



Login with Nikhef SSO and our brand-new OIDC provider … under test …

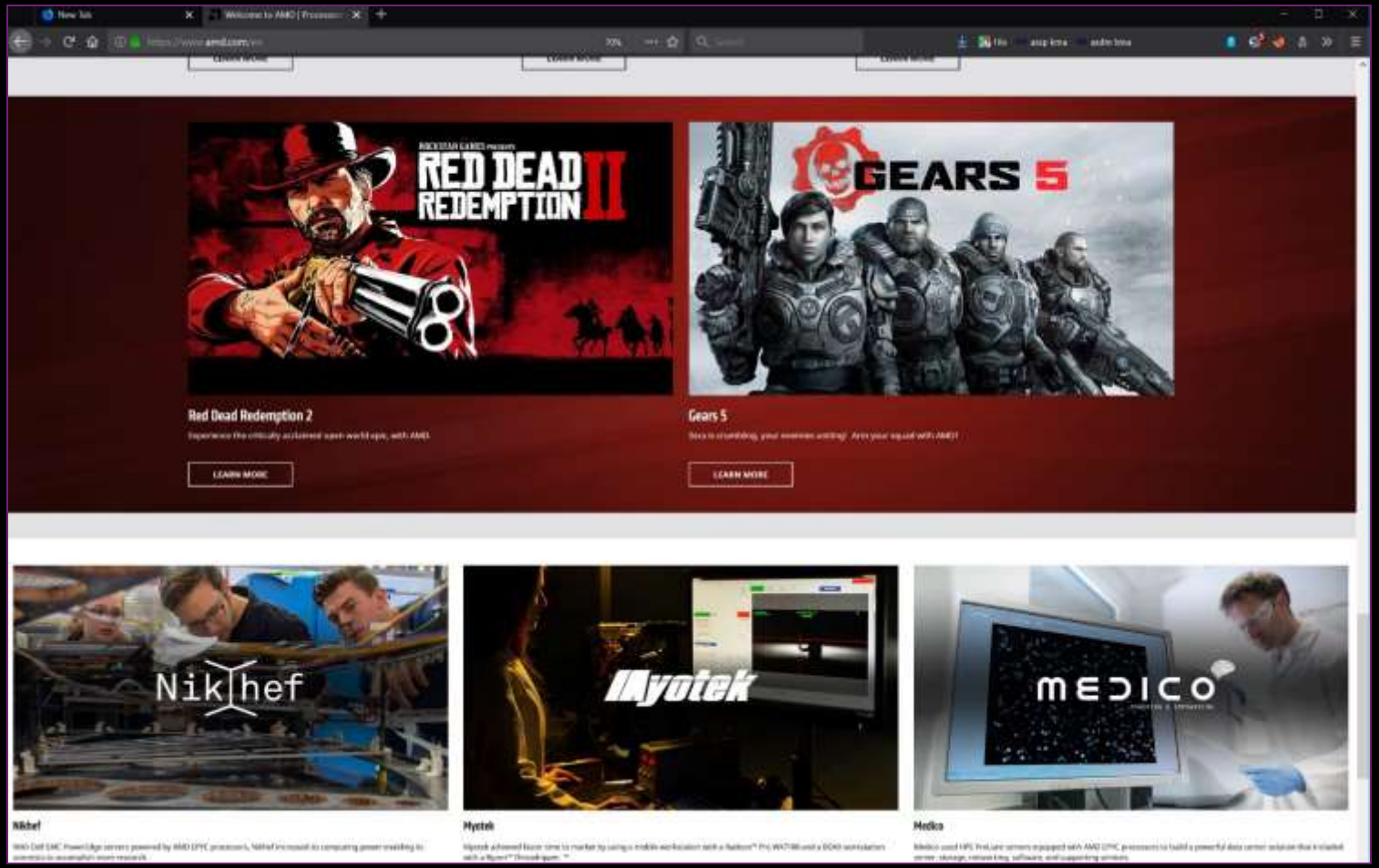
PARTLY IT'S 'JUST' HARDWARE

Old clusters get re-used nowadays for our experimental cloud service

But what will be the new 'best system' varies:

- some jobs require big mem (8GB/core)
- others are happy with many cores
- on-board throughput often the bottleneck

And most effective and efficient design changes continuously …


RED DEAD REDEMPTION II
Red Dead Redemption 2
LEARN MORE
GEARS 5
Gears 5
LEARN MORE
Nikhef
Nikhef
Mycelium
Mycelium
MEDICO
Medico

BIGGER IS BETTER - IF YOU KEEP IT TOGETHER

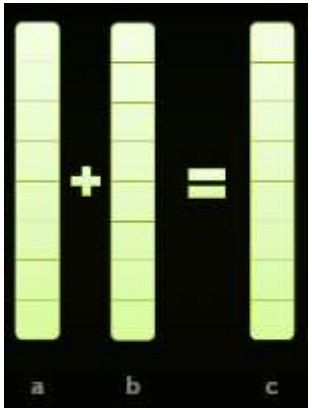
Common element: moving data is ‘expensive’, so:
keep on computing as long as you can, and don’t move data around

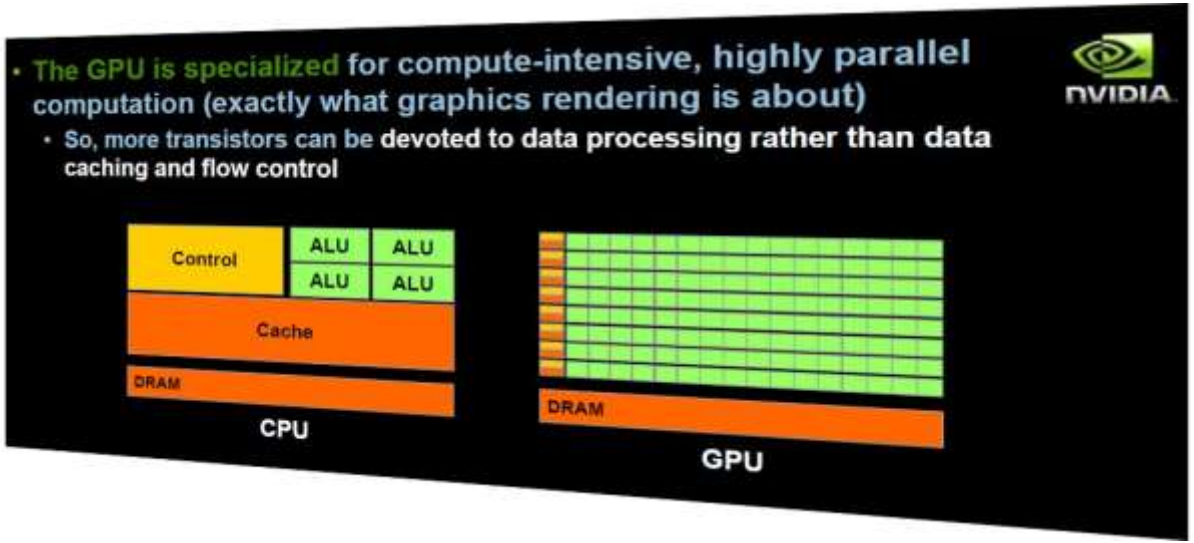
- AMD single-socket many core system (now with NVMe scratch space)
  no useless cache coherency delays, more direct memory access
  *and **a lot** of PCI-e lanes to get data from net through storage to CPU*

- similarly, keep your GPU cores busy … by making the problem larger!

ONCE YOU HAVE DATA: COMPUTE, DON'T ARGUE

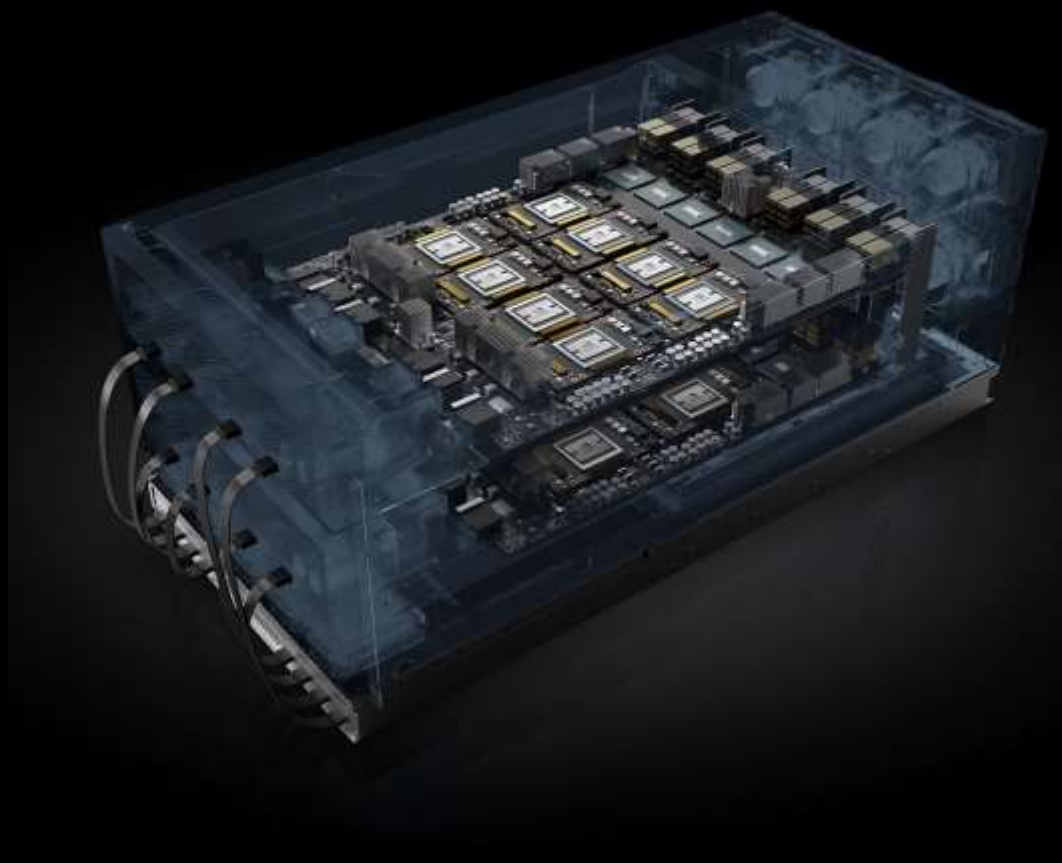
moving to GPGPUs





GPUs are great for performing the
Same operation (Instruction) on
Multiple Data elements (SIMD)
but that data has to be there, and shipped there …

Sources: NVidia 'Massively Parallel Computing with CUDA' and https://www.nikhef.nl/~janjust/presentations/PDP-groepsoverleg-20140423jjk.odp



HGX-2
2 petaFLOPS tensor operations 250 teraFLOPS single-precision 125 teraFLOPS double-precision
16x NVIDIA Tesla V100
512GB total 16TB/s bandwidth
81,920
10,240
NVSwitch powered by NVLink 2.4TB/s bisection bandwidth

Source: datasheets and images from
https://www.nvidia.com/en-us/data-center/hgx/

PRECISELY?

2 petaFLOPS tensor operations

250 teraFLOPS single-precision

125 teraFLOPS double-precision

PREPARE FOR OPEN GPU PLATFORMS

World is changing .. rapidly:

- the time of ‘GPU is affordable’ was killed by nVidia
- for efficient computing, we need competition – which is there

Going beyond just CUDA allows you (and our system architect) to get best now-current hardware and performance-price point

We can ‘transpile’ … but why not use generic languages

- both are C++ derivatives: CUDA, ROCm, …
- *e.g.* HIP: C++ Heterogeneous-Compute Interface for Portability

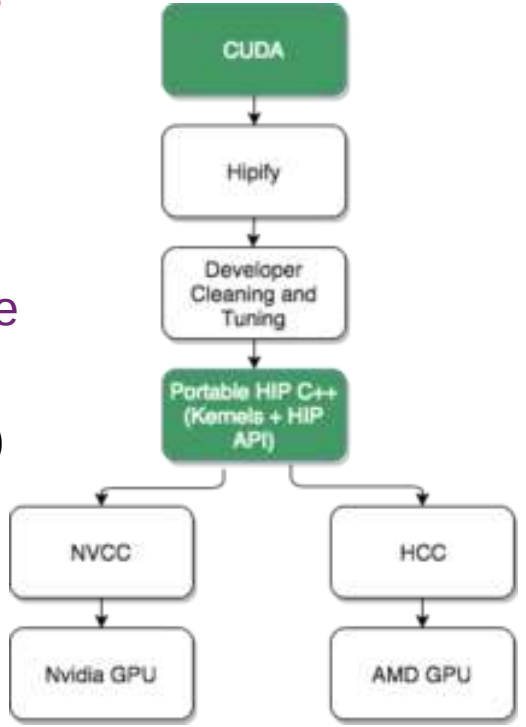


Image: https://towardsdatascience.com/on-the-state-of-deep-learning-outside-of-cudas-walled-garden-d88c8bbb4342


x8/x16 Riser 3 Cabling
Connects to Riser 3 Power Connector on MB
Connects to Riser 3 Sideband Connector on MB
Connects to PCIe Connector 2 on MB
Connects to PCIe Connector 3 on MB

AND IF IT DOESN'T QUITE FIT …

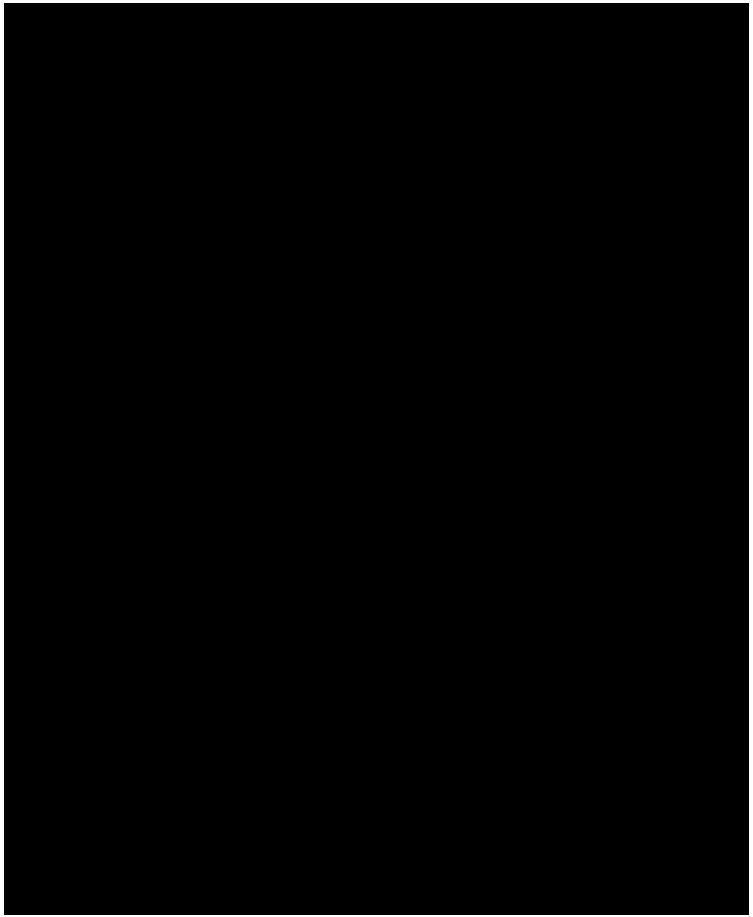
SuperMicro (branded as Lambda Blade)
4U chassis, supporting 10 consumer-grade GPUs …
… with a bump

Image source: https://lambdalabs.com/products/blade

THIS IS ALSO JUST REALLY BIG (BUT NOT THAT FAST …)

We have now sufficient fast storage
to serve the cluster compute jobs

so: we can add ‘sudder’ storage and
introduce tiered (hierarchical) storage



SINCE WE NOW HAVE ENOUGH STUFF THAT'S FAST …

We know and love fast storage

12 MiB/s/TiB - about 4 PiB configured (both DNI/NL-T1 and Stoomboot)

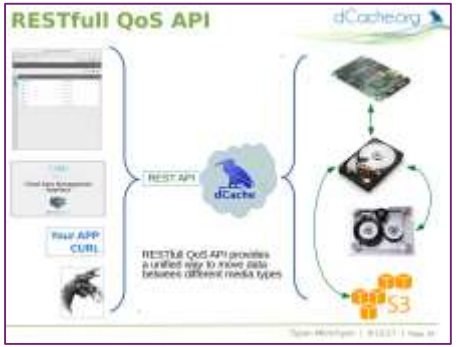
'hooikanon': 240 (4x60) spindles, 12 TByte disks, 4x100Gbps network
4x IBM SL922 Power9 PPC servers, 8x NetApp E5700 controllers, 4 trays

TIERED STORAGE ENTAILS SOME COORDINATION

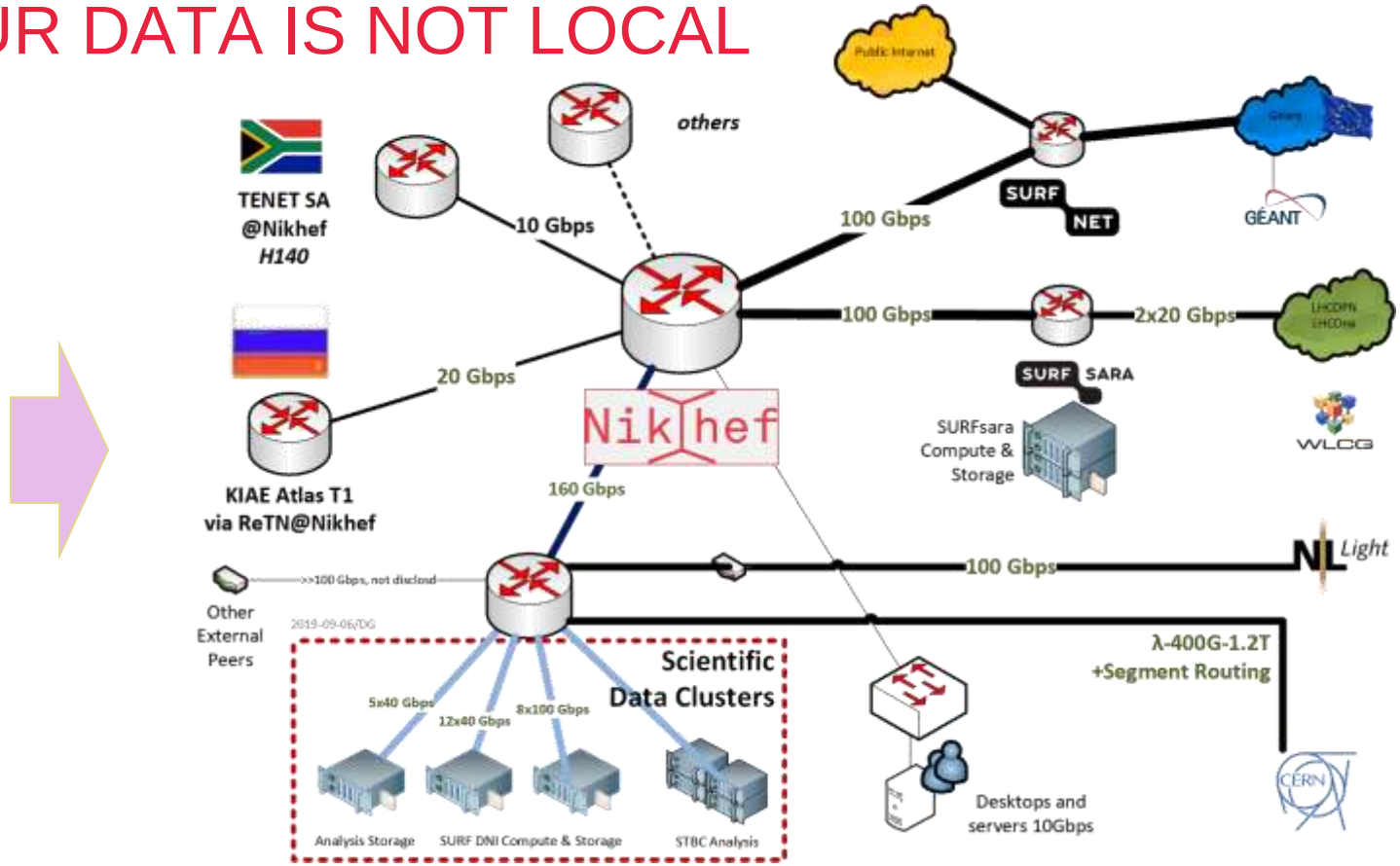
Fast
- dCache STBC today
- Grid storage

Bulk
- ‘/data’
- dCache bulk tier

Controlled
- NL-T1 tape tiers
- SURF Data Repository



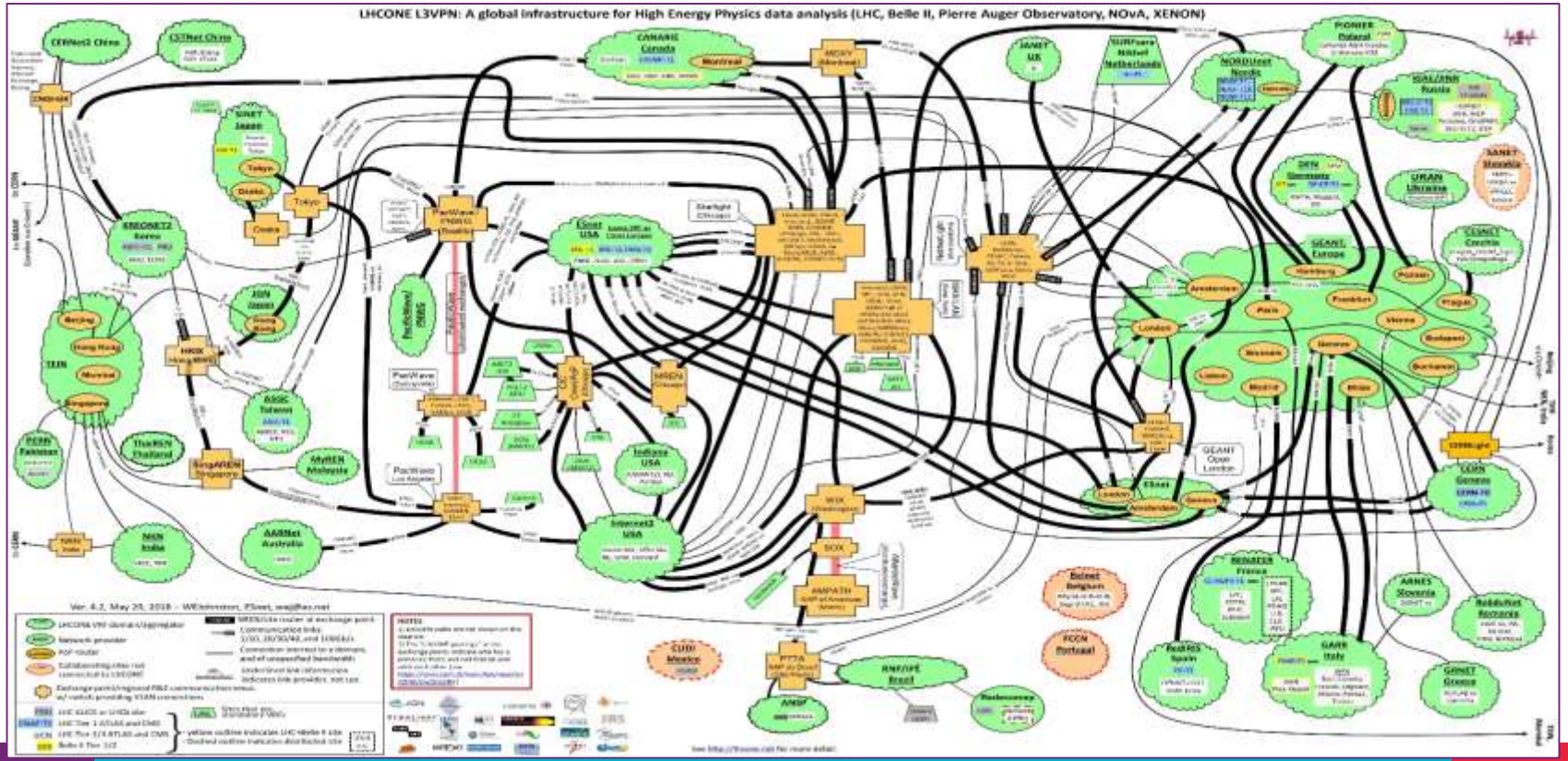
AND IF YOUR DATA IS NOT LOCAL



Small is beautiful ... or is it ?

Left: IBR-LAN (1996) in H1.40; Right: Nikhef peering visualisation (Dec 2019)

LHC OPEN NETWORKING ENVIRONMENT (LHCone)



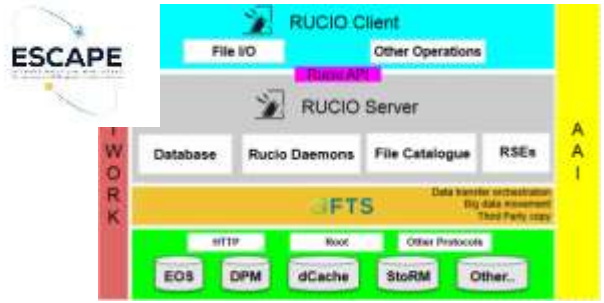
Small is beautiful ... or is it ?

Image: Bill Johnston et al.

SMALL PACKETS – BIG TROUBLE? A 1 **B**PPS SYSTEM

But long-distance fat networks work fine for ‘big files’

- *bandwidth* x *delay* hampers remote random access
- which is why ESCAPE develops data lake caching



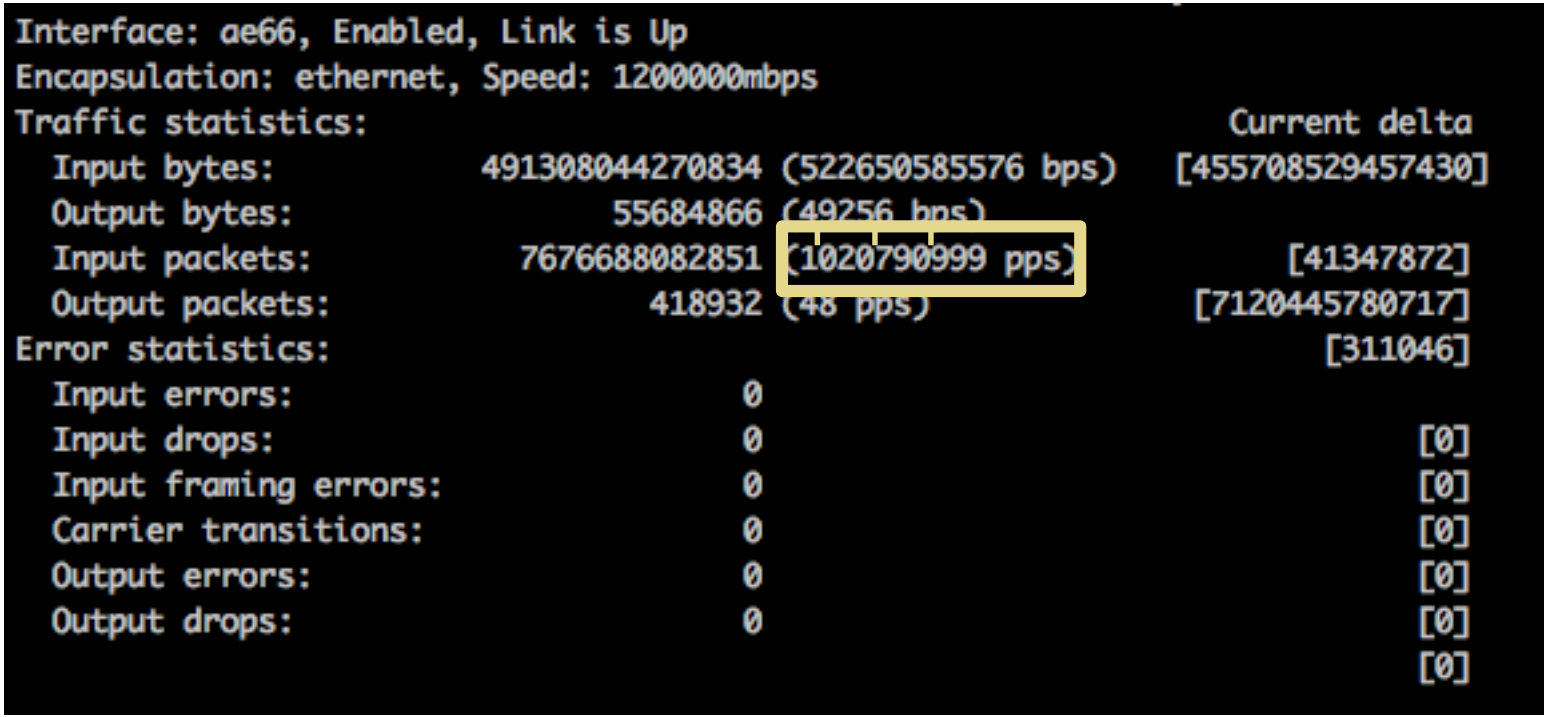
Some data is inherently ‘small’ and packetized

- ‘telemetry’ data
- remote distributed event processing

but sending many packets is far more challenging on network ASIC design – so let’s test it!

Graphic Data Lake, Simone Campana, ESCAPE WP2 DIOS
Image 1Bpps machine by Tristan ☺

OUR WORLD RECORD: 1 BILLION PACKETS PER SEC



14-06-2019: 1 billion pps *i.e.* 1 Gpps (and 522Gbps)

https://wiki.nikhef.nl/grid/1Bpps_Machine

Image: Tristan Suerink

Nikhef

BUT WHAT IS DATA TO US …

Image: Tristan Suerink

WE CAN'T KEEP HIDING SYSTEM ARCHITECTURE

We can cover some but not all complexity for you – so to build the next generation efficient and effective architectures, we *want you to try it and co-evolve*

but **experimental services** are **not production services** like STBC of dCache
please treat them as such – they may be useful, they may break, or go away later

- GPU systems – we have reference systems (plofkip, ballenbak) and will grow the STBC-GPU (production) section
- Early cloud and jupyter hub
- Global networks and peerings
- *and of course 'odd' services like a Vidyo phone bridge, XOA cloud, NikhefTV, &c*

'if it breaks, you get to hold the pieces' – but tell us so we can evolve the design

BUT NOT THIS KIND OF EXPERIMENTAL PLEASE …

ALTHOUGH …

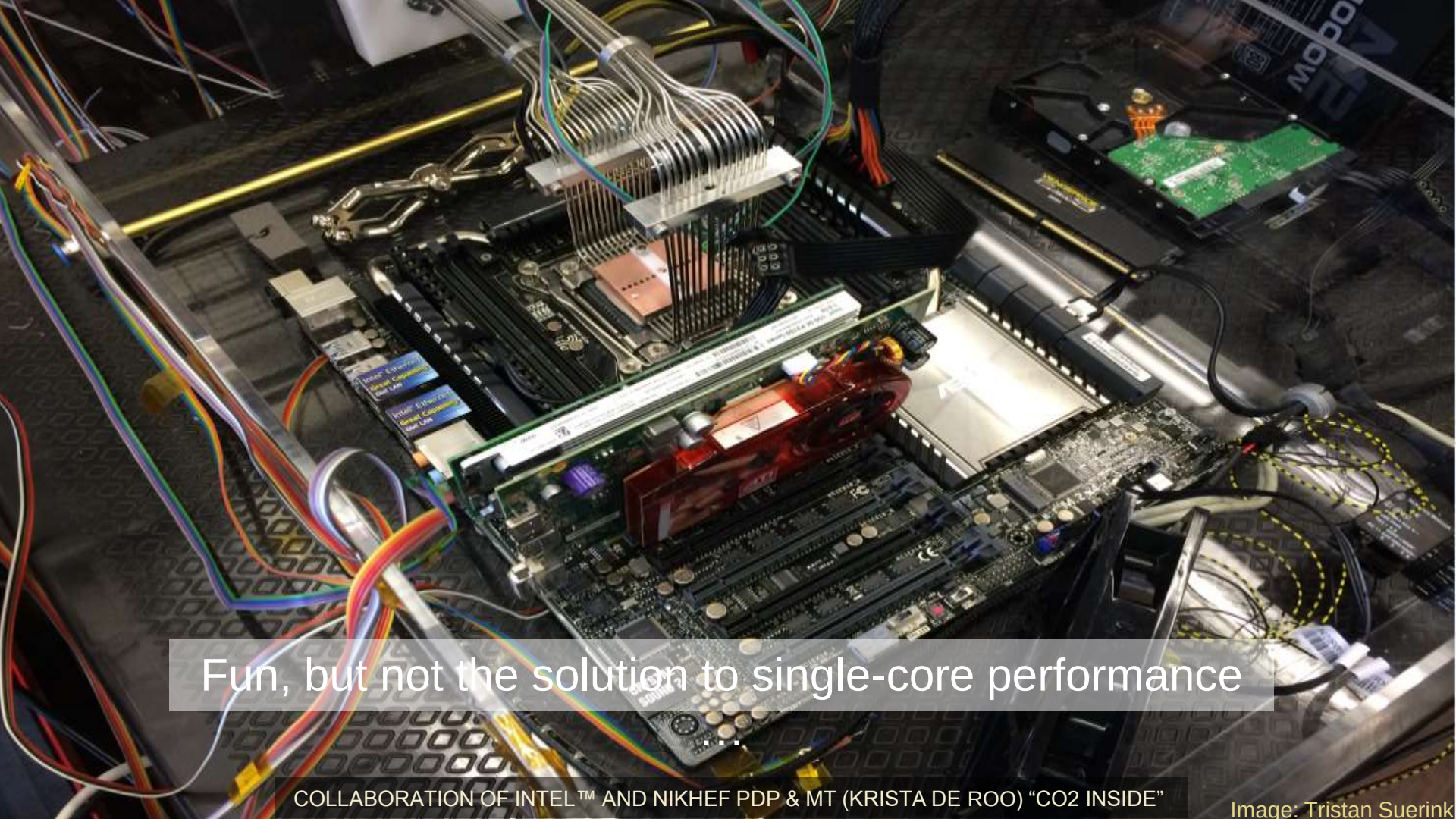
Image: Tristan Suerink

BIGGER IS OFTEN BETTER!

But our new systems architectures require **joint action at every level**
the only way we can grow to the scales needed

- Hierarchal storage: organize analysis or ‘classify’ data heaps
- GPU computing: most efficient vector processing requires you to think big – and across platforms
  we can get much more GPU power if you don’t stick to one platform …
- Networks for event-sized data flows

Enjoy experimental services and help us, our DNI, and our community to scale and grow – but be prepared to change and anticipate it breaking



Fun, but not the solution to single-core performance

COLLABORATION OF INTEL™ AND NIKHEF PDP & MT (KRISTA DE ROO) “CO2 INSIDE”

Image: Tristan Suerink

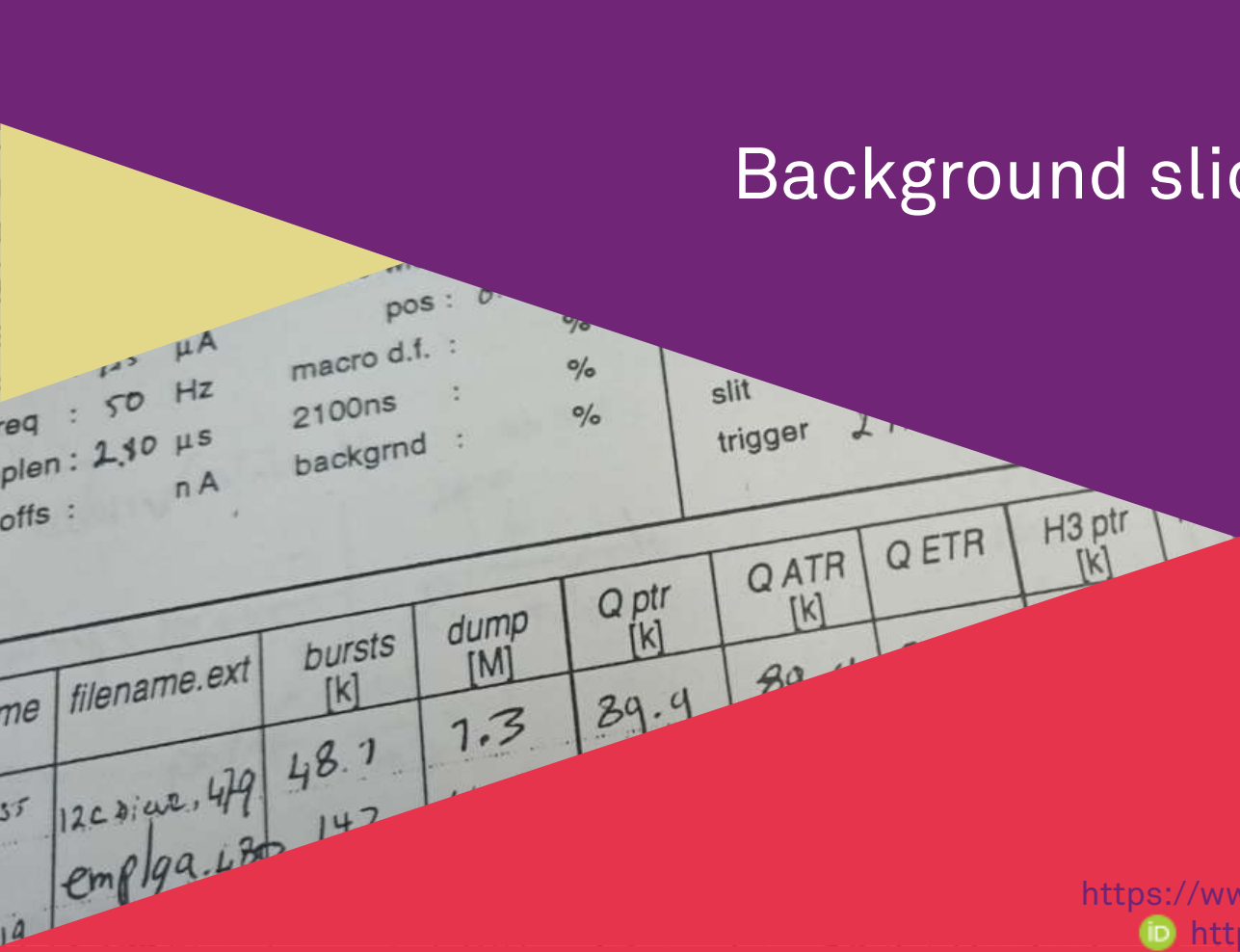
Note: some foils contain confidential or commercially sensitive information
*Please **do not share** beyond Nikhef without asking*

‘Size Helps!’

David Groep
davidg@nikhef.nl
https://www.nikhef.nl/~davidg/presentations/
https://orcid.org/0000-0003-1026-6606

SURF
part of the work is co-supported by SURFsara under the Dutch National e-Infrastructure innovation programme

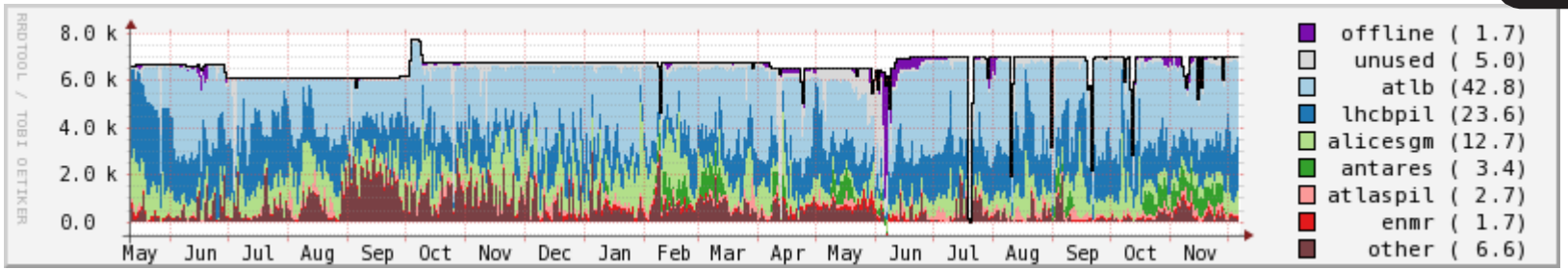
Background slides

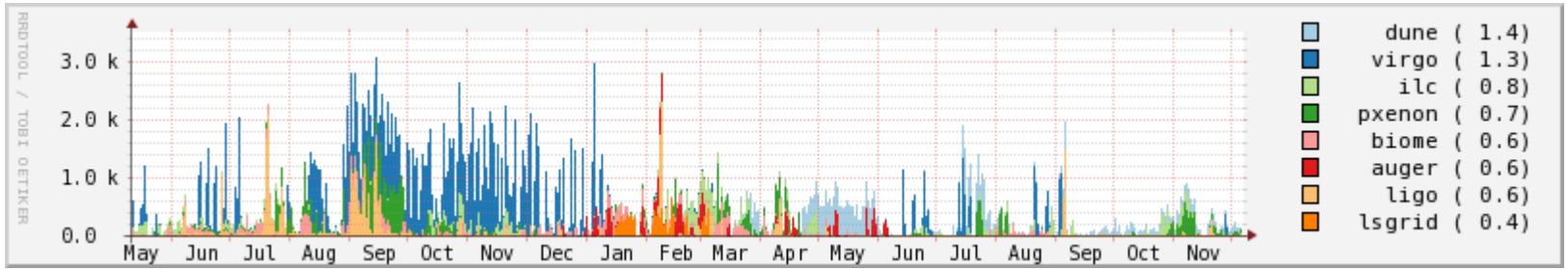


David Groep
davidg@nikhef.nl
https://www.nikhef.nl/~davidg/presentations/
https://orcid.org/0000-0003-1026-6606

WE STILL HAVE ENOUGH FOR NEW SERVICES …

More communities emerging - including the 1800 cores supplied to SURF





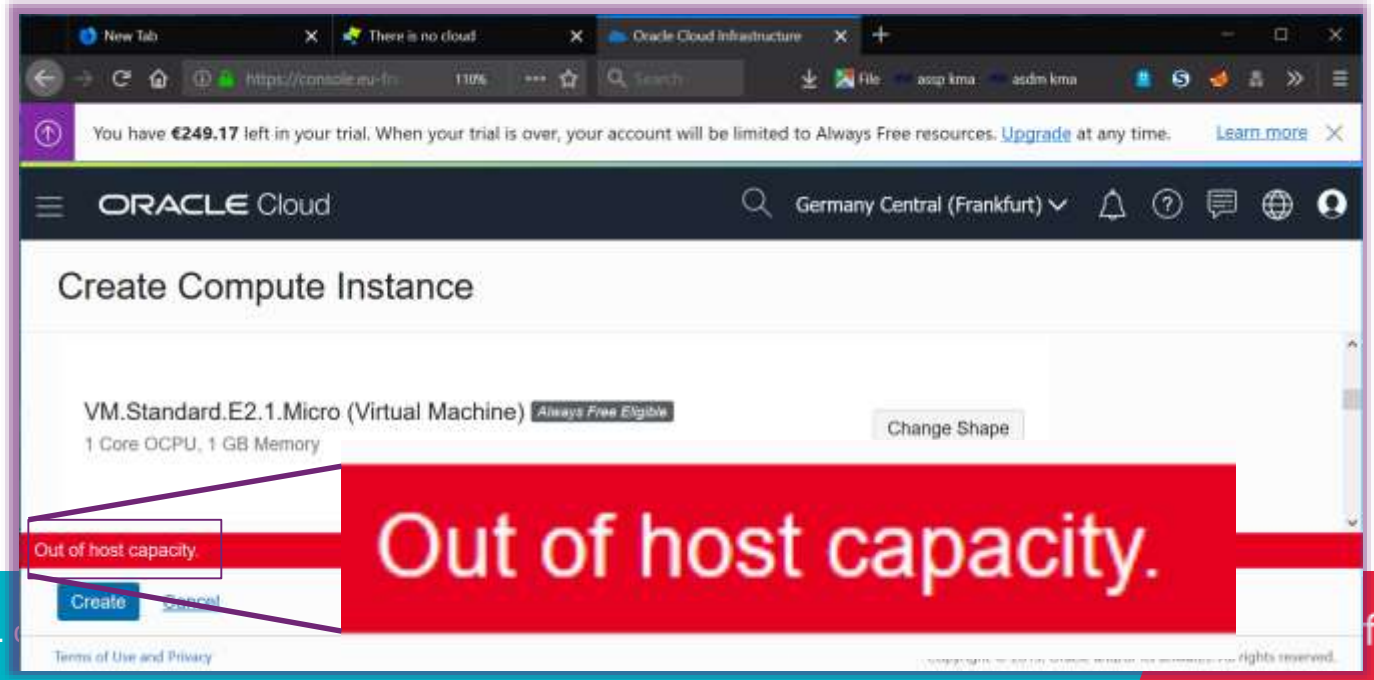
‘CLOUD’ WORKS BECAUSE OF SCALE

The impression of elasticity comes with overprovisioning of resources

… and thus at a cost for the ecosystem as a whole

… ‘the best cloud is both full and empty at the same time’

we now have sufficient size to be ‘creative’ and designate ‘trial’ resources



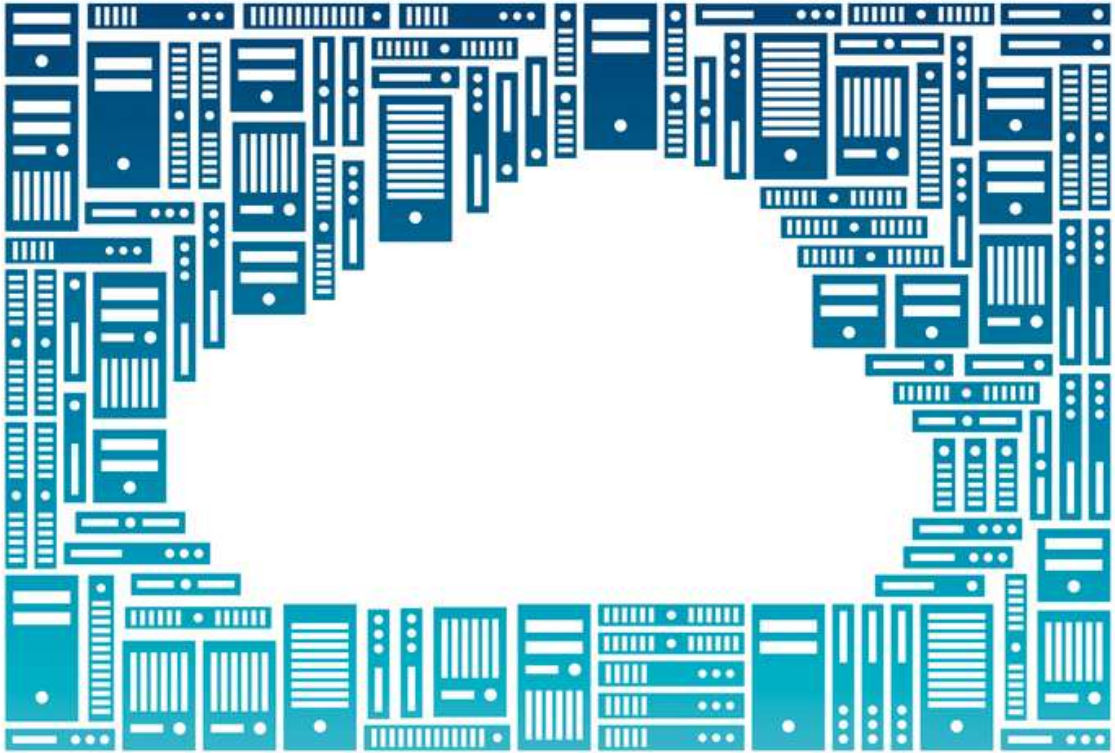
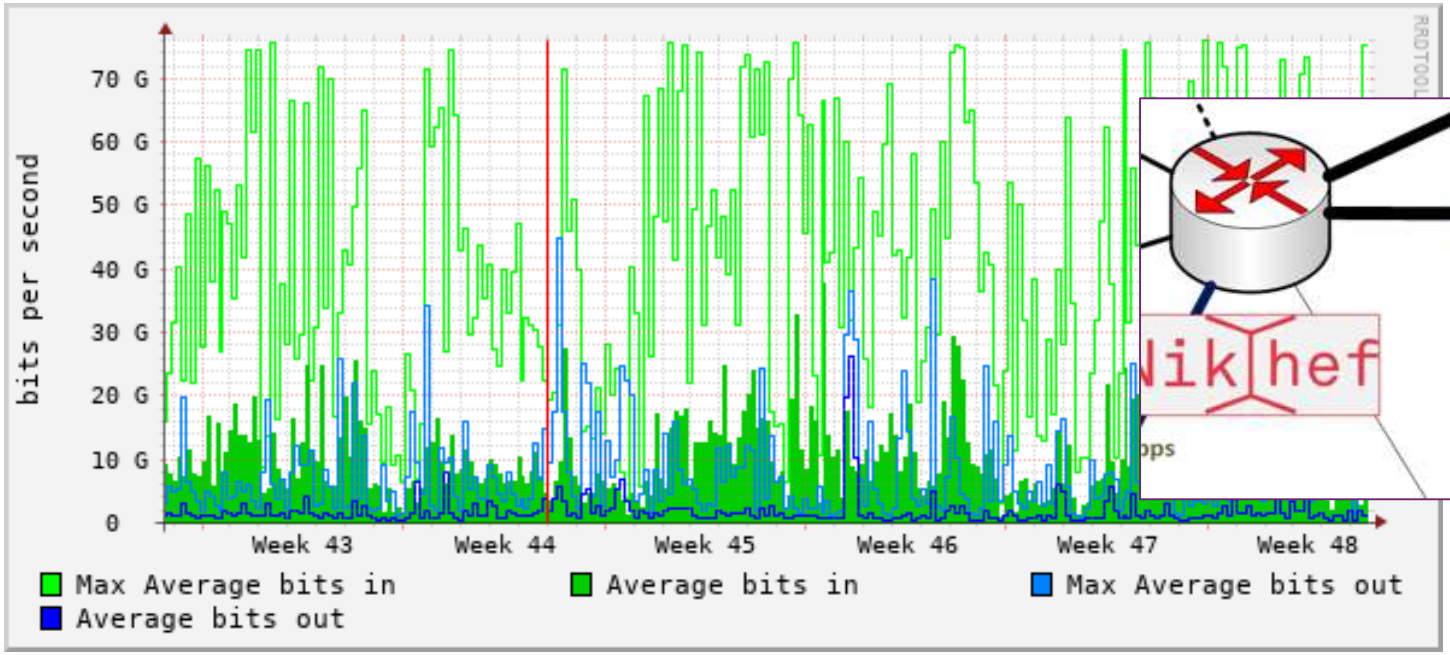
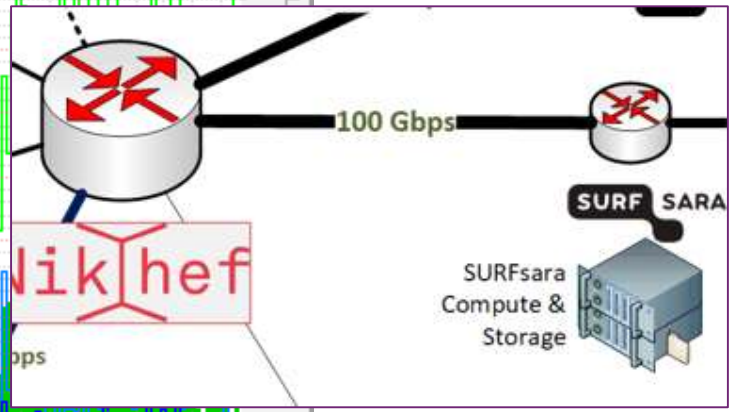


Image source: Free Software Foundation Europe - https://fsfe.org/

YOU GET LOTS OF DATA FROM SURFSARA

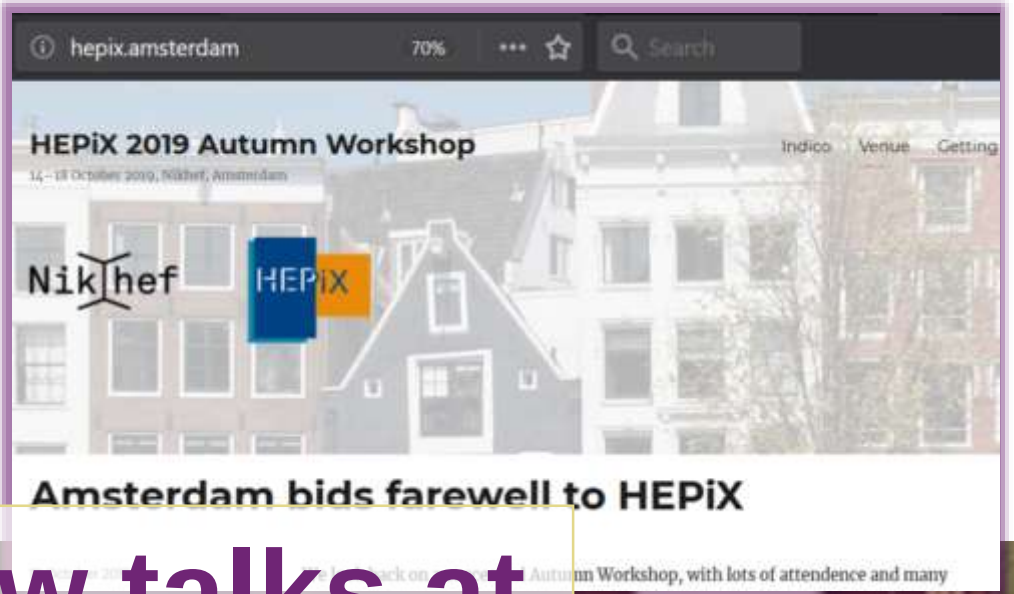




80Gbps SURFsara cap, observed November 2019
http://cricket.nikhef.nl/cgi-bin/grapher.cgi?target=%2Fparkwachter.ipmi.nikhef.nl%2Fet-9_3_0

HEPIX2019 – MANY SCALES ALL AT ONCE

- Next-gen processors and GPGPUs
- Disk, tape, data lake, data preservation
- Condor, ARC, Cloud, or Slate?
- Networks, and LHCOne beyond LHC
- …



Review talks at tv.nikhef.nl