Convergence of computation and data workflows

IS-ENES Workshop on Workflows and Metadata Generation
Lisbon, PORTUGAL

V. Balaji

NOAA/GFDL and Princeton University

28 September 2016

Amy Langenhorst 1977-2016

Principal Developer of the FMS Runtime Environment FRE.

Outline

Outline

Power-8 with NVLink

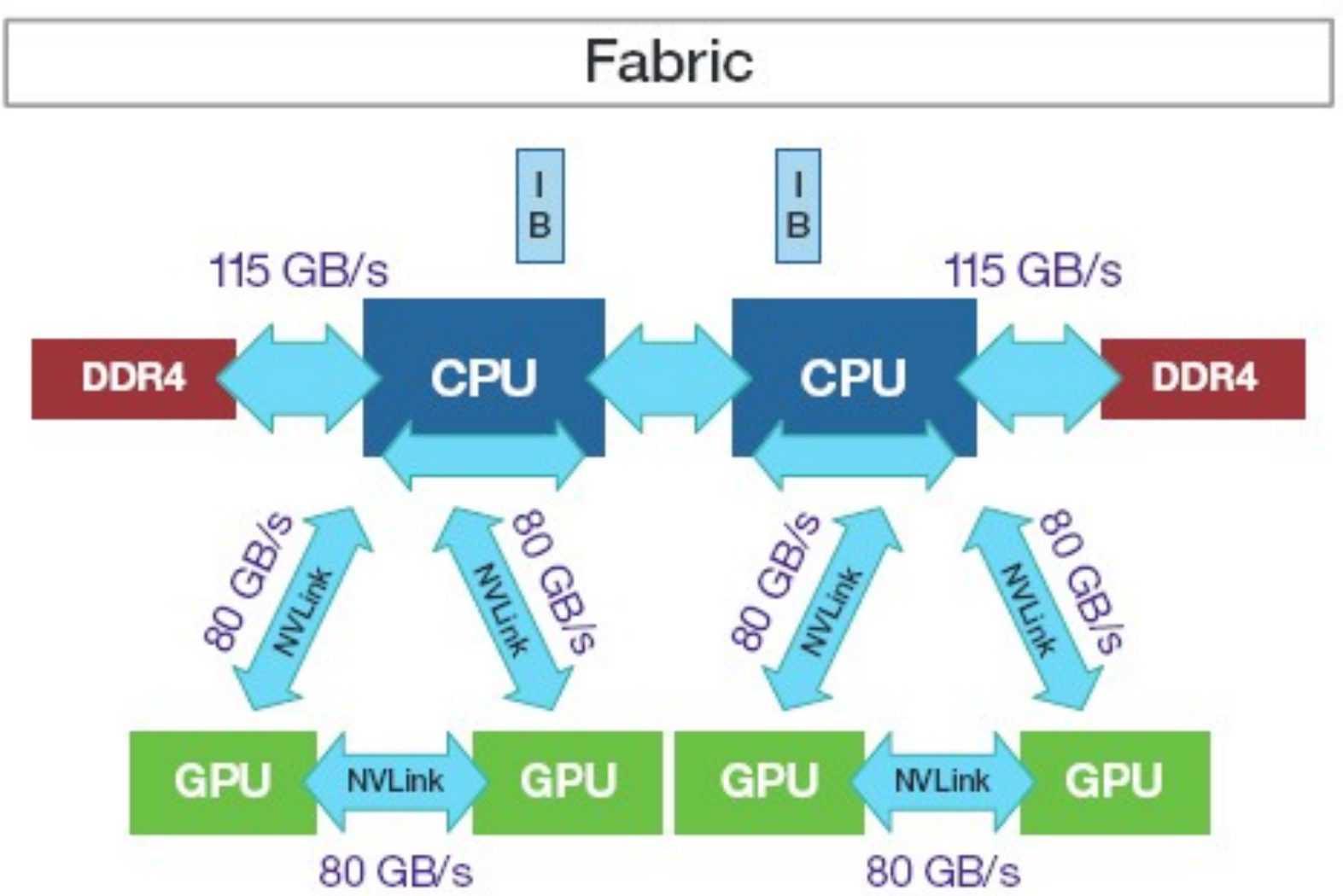


Figure courtesy IBM.

KNL Overview

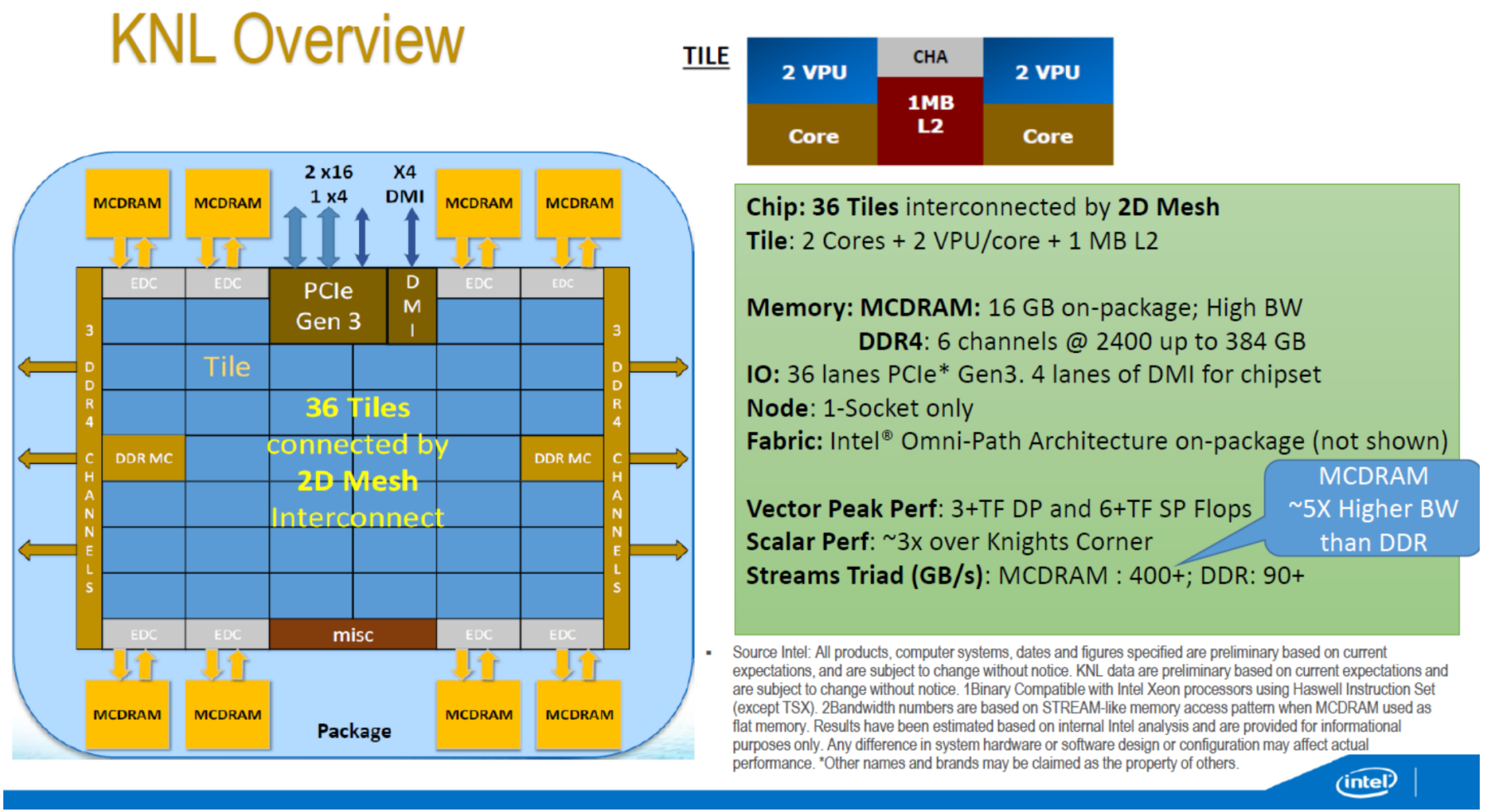
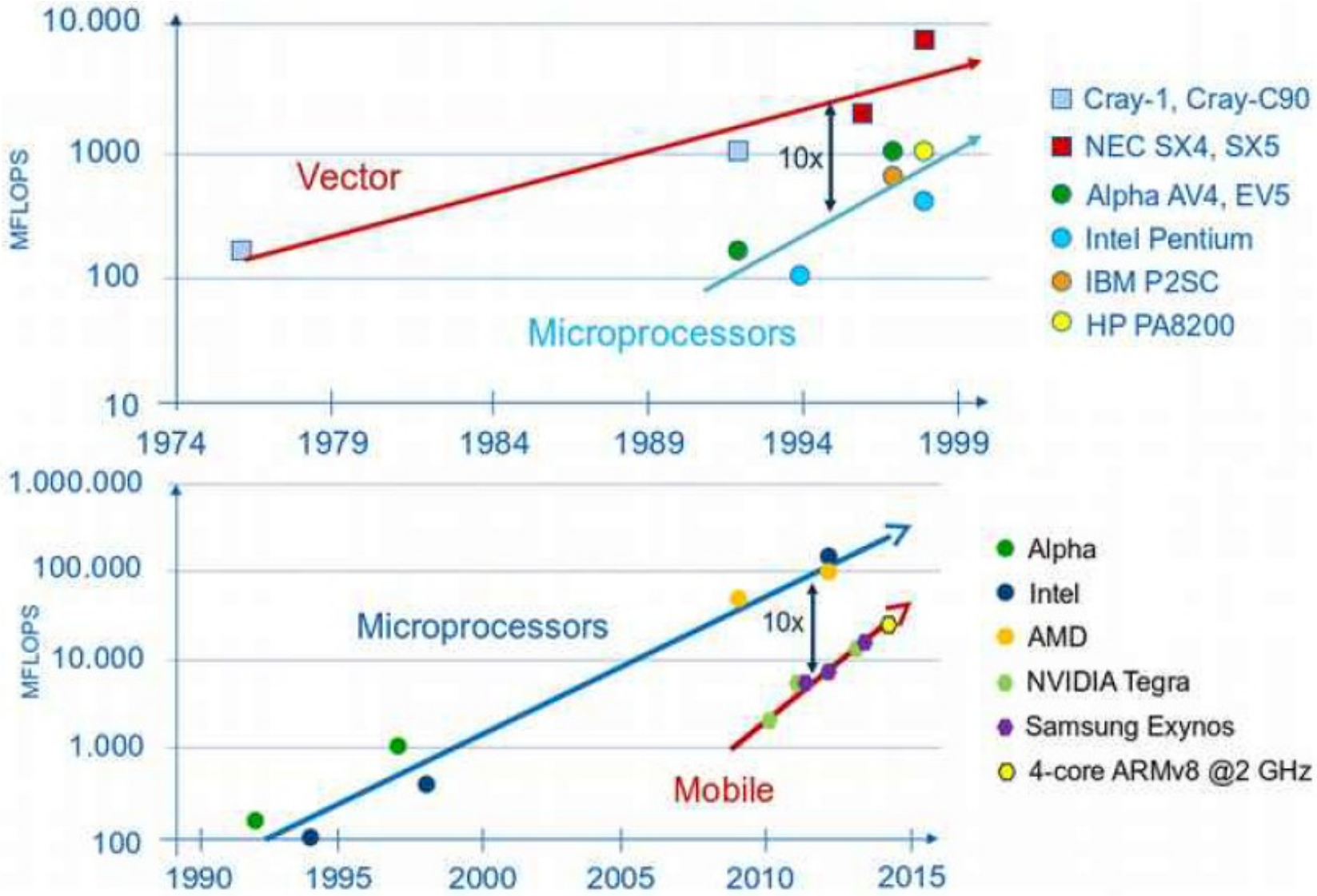


Figure courtesy Intel.

The inexorable triumph of commodity computing



... means ARM? From *The Platform*, Hemsoth (2015).

Irreproducible Computing, Inexact Hardware

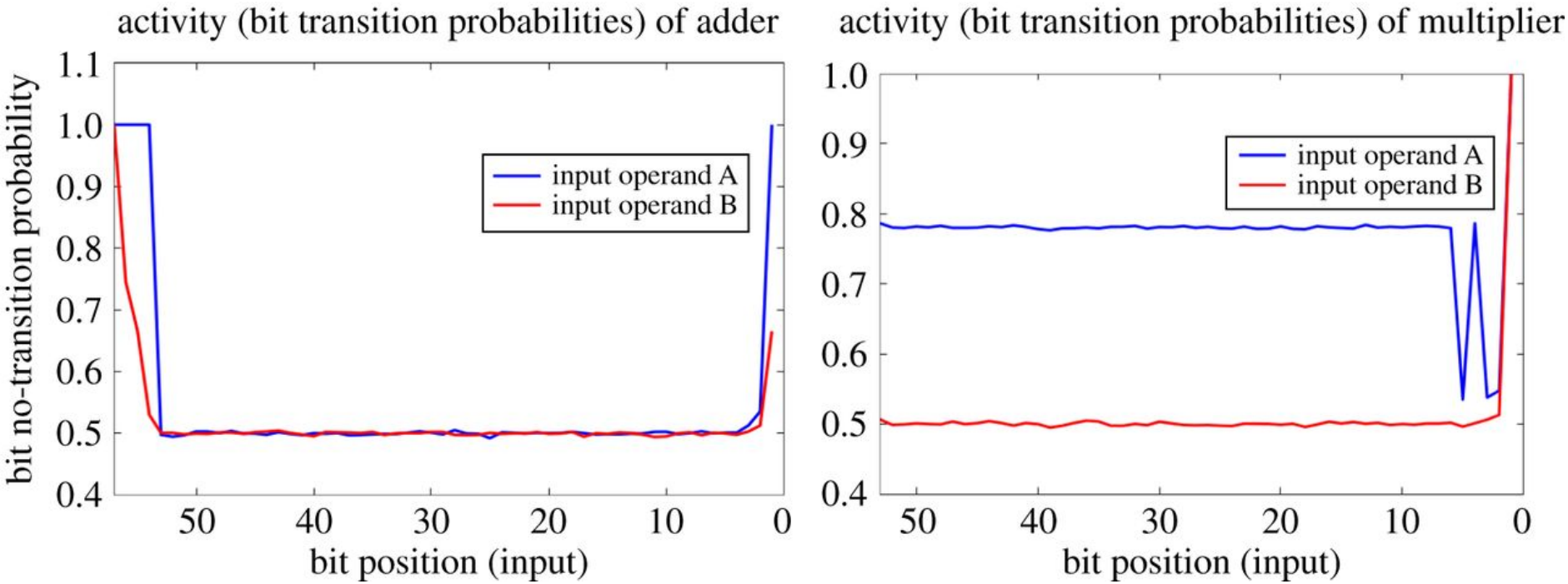


Figure 1 from Düben et al, *Phil. Trans. A*, 2016. Which bits can we allow to be "inexactly" flipped? Lorenz 96 as canonical test case of non-linearity and chaos.

Irreproducible Computing, Inexact Hardware

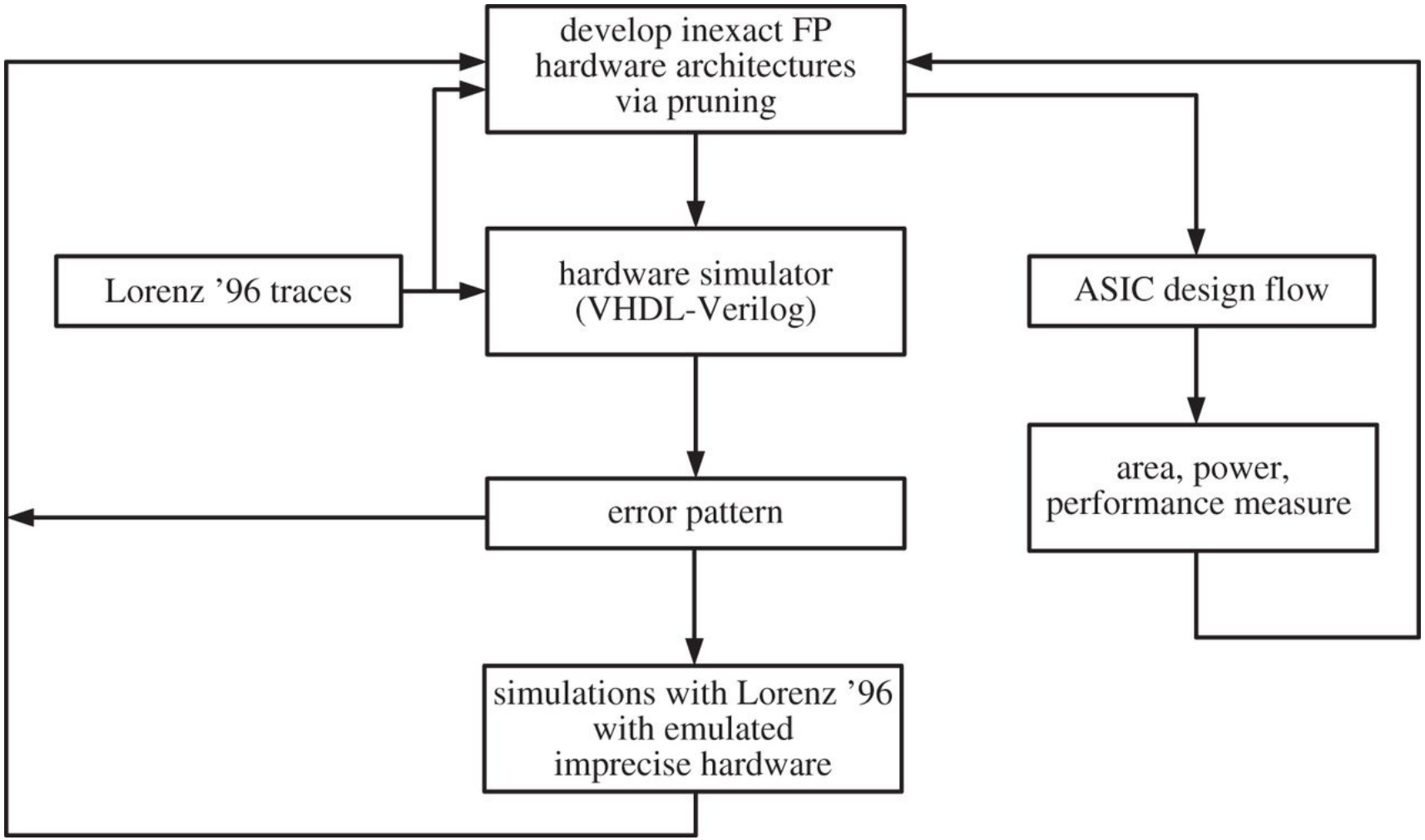


Figure 2 from Düben et al, *Phil. Trans. A*, 2016.

COSMO: energy to solution

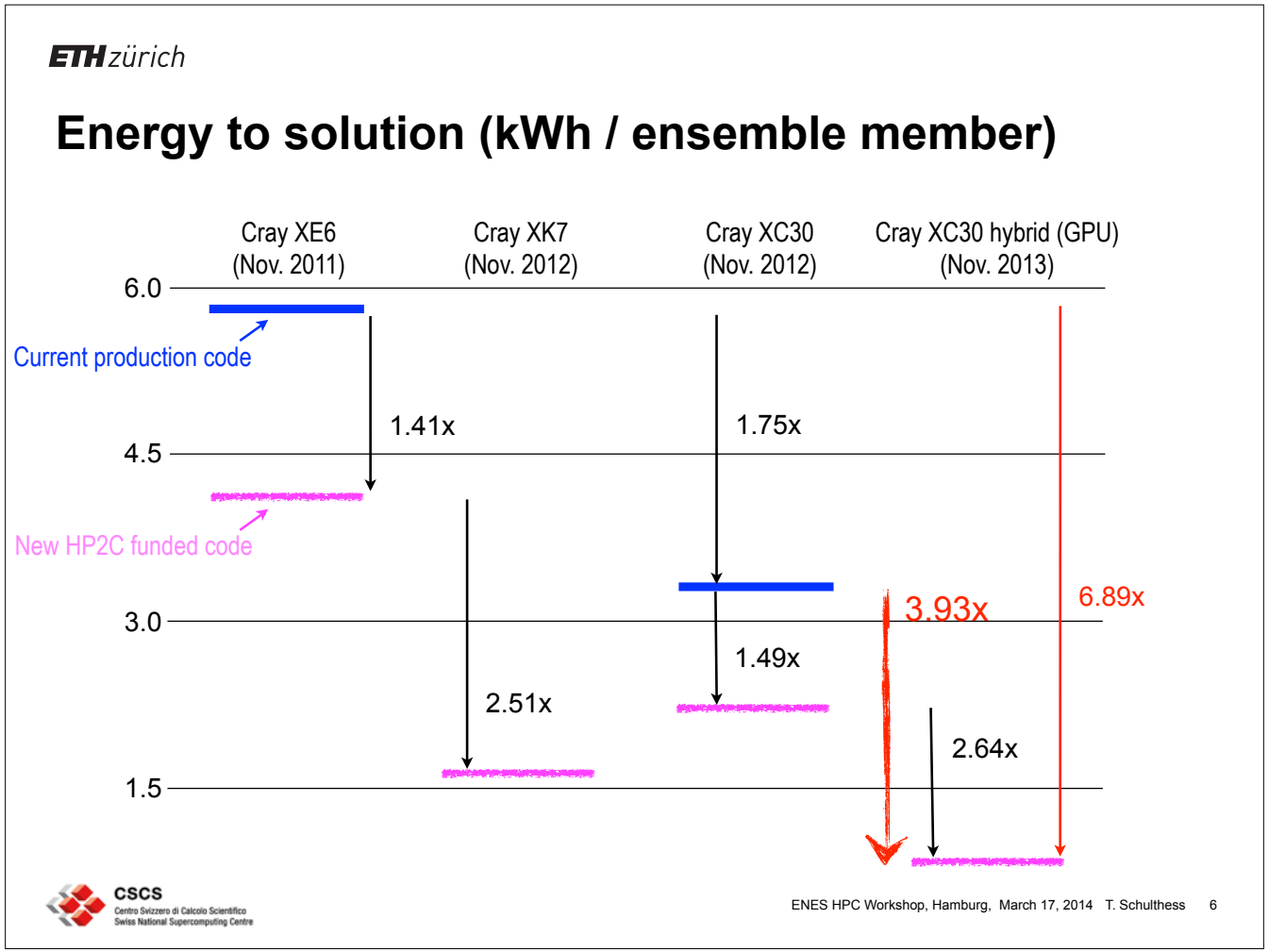


Figure courtesy Thomas Schulthess, CSCS.

JPSY comparison across ESMs

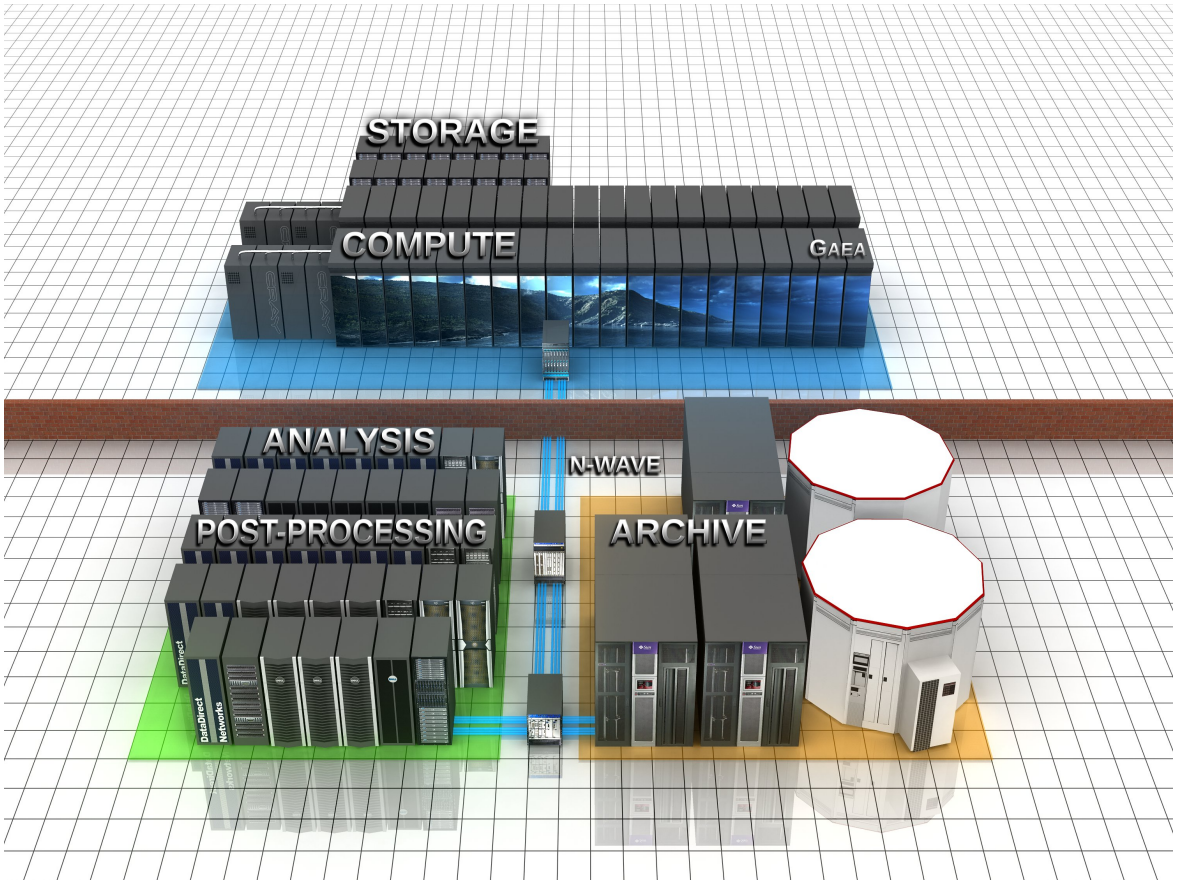
Model	Machine	Resol	SYPD	CHSY	JPSY
CM4	gaea/c2	1.2×10^8	4.5	16000	8.92×10^8
CM4	gaea/c3	1.2×10^8	10	7000	3.40×10^8

- Comparative measures of capability (SYPD), capacity (CHSY), and energy cost (JPSY) per "unit of science".
- Can you have codes that are "slower but greener"? Algorithms that are "less accurate but more eco-friendly"?
- From Balaji et al (2016), in review at GMDD. **http://goo.gl/Nj1c2N**

Workflows for the exascale

- Billion-way concurrency still a daunting challenge for everyone: no magic bullets anywhere to be found.
- Exotic hardware is on the way; this is quite likely the last generation of conventional hardware. Computing is likely to become irreproducible.
- Software investment paid back in power savings (Schulthess). Energy to solution will become key metric.
- More threading needs to be found: to fit 10^{18} op/s within a 1 MW power budget, an operation should be 1 pJ: data movement is $\sim$10 pJ to main memory; $\sim$100 pJ on network!
- DARPA: commodity improvements will slow to a trickle within 10 years: go back to specialized computing?
- DOE: double investment in exascale.

A network of compute and data nodes



FRE and other elements in the GFDL modeling environment manage the complex scheduling of jobs across a distributed computing resource.

... a global network of compute and data nodes

Workflow task is to minimize data flow across the global network.
Figure courtesy IPSL.

Outline

Examples of DAG parallelism

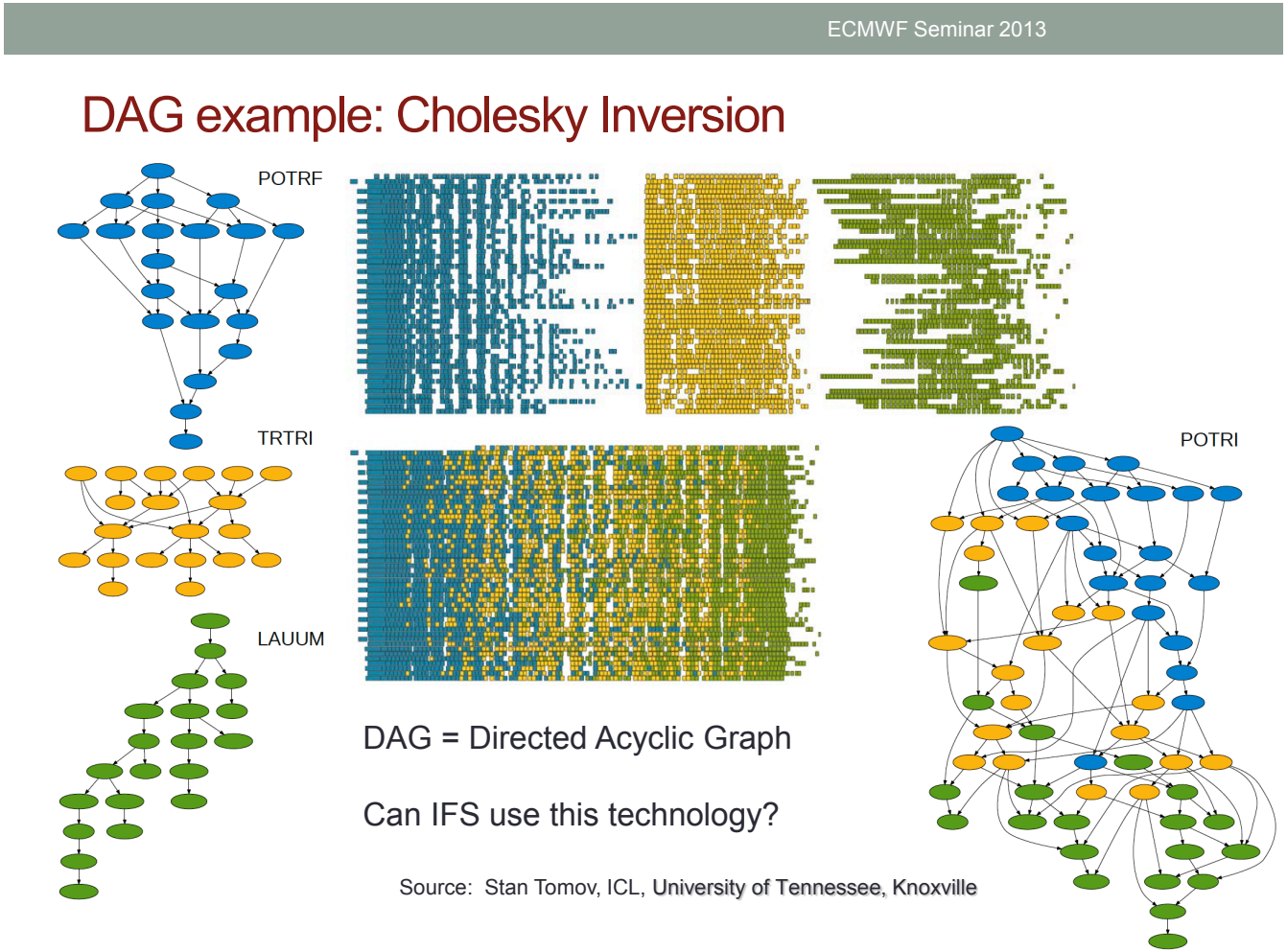
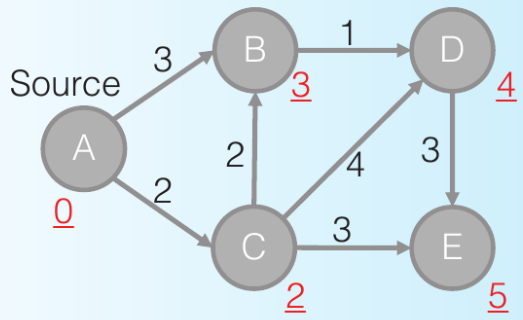


Figure courtesy George Mozdzynski, ECMWF.

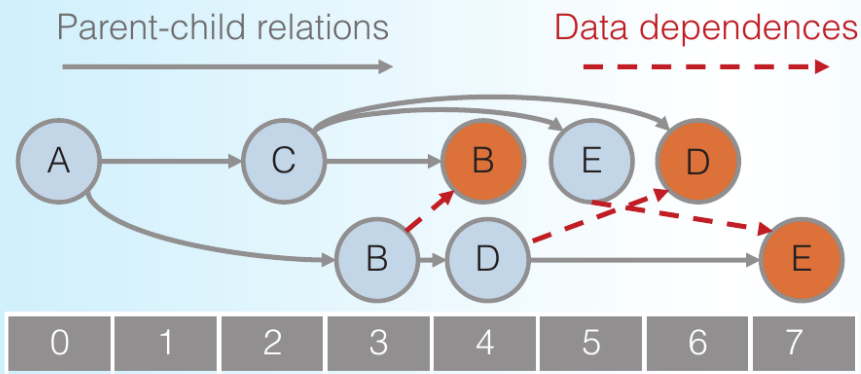
SWARM for DAGs

```
prioQueue.enqueue(source, 0)
while prioQueue not empty:
    (node, dist) = prioQueue.dequeueMin()
    if node.distance not set:
        node.distance = dist
        for nbr in node.neighbors:
            d = dist + edgeWeight(node, nbr)
            prioQueue.enqueue(nbr, d)
    else: // node already visited, skip
```

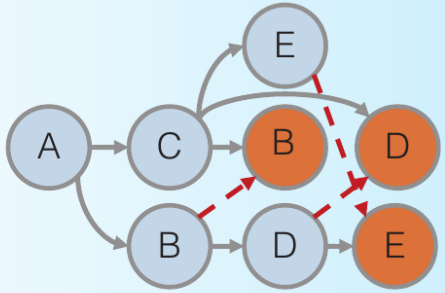
(a)



(b)



(c) Order = Distance from source node



(d)

Jeffrey et al, *IEEE Micro* 2016.

KNL Overview

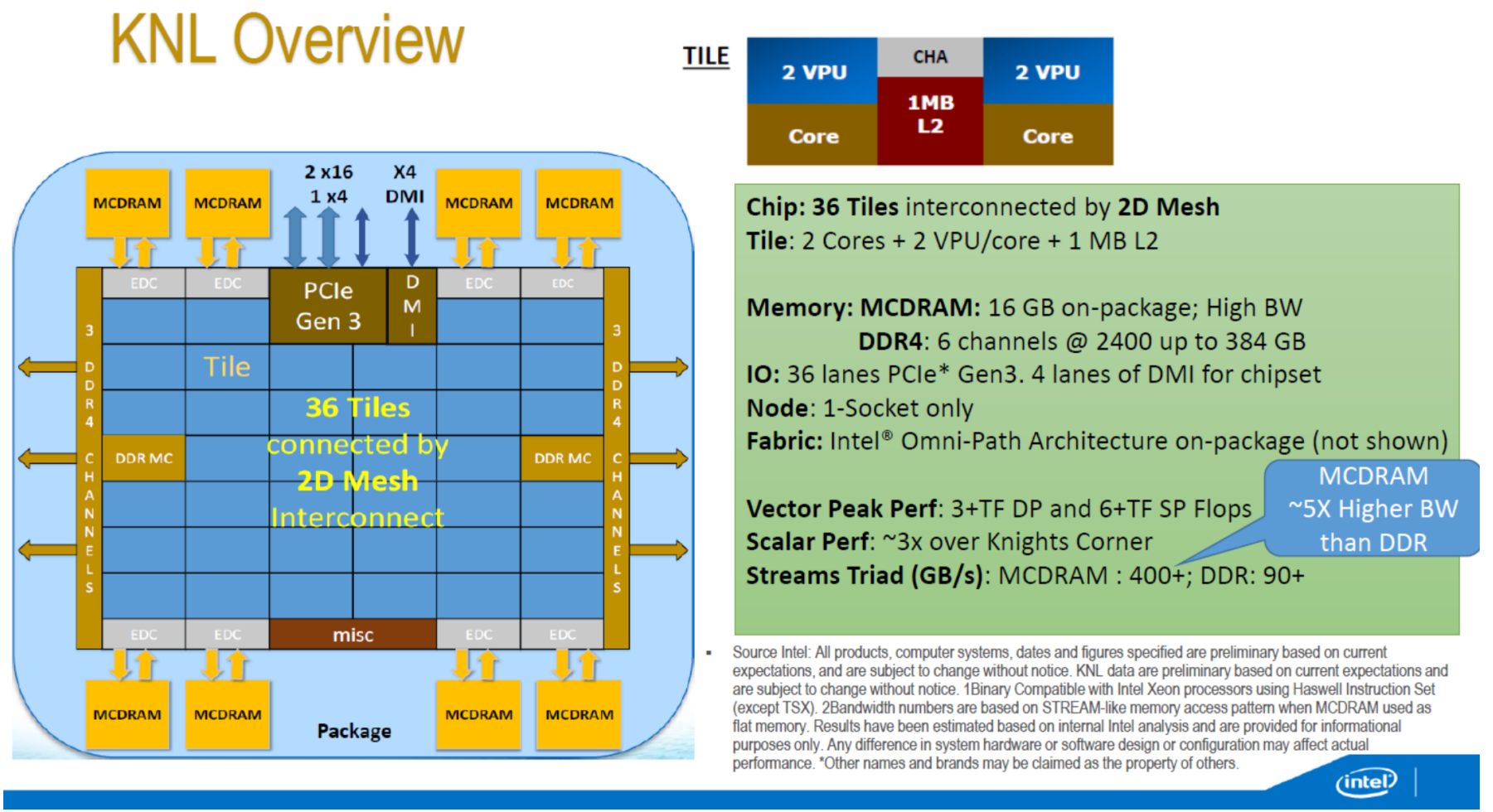
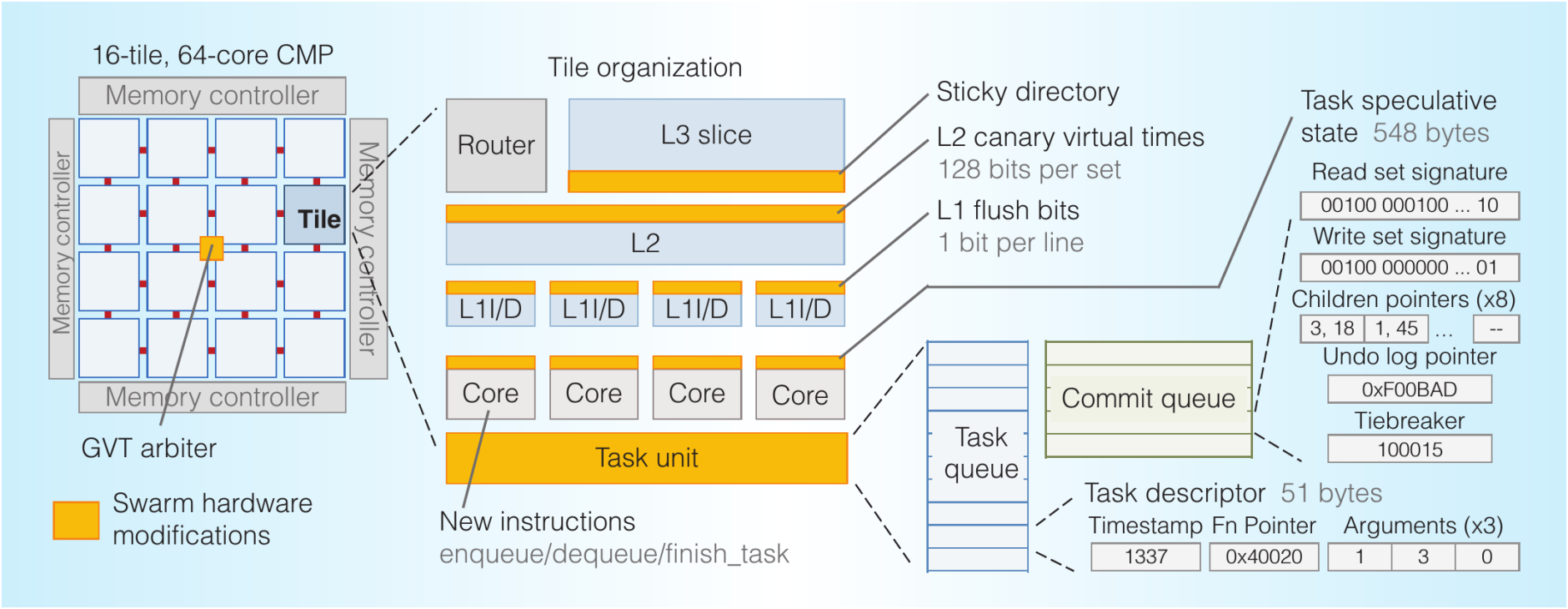


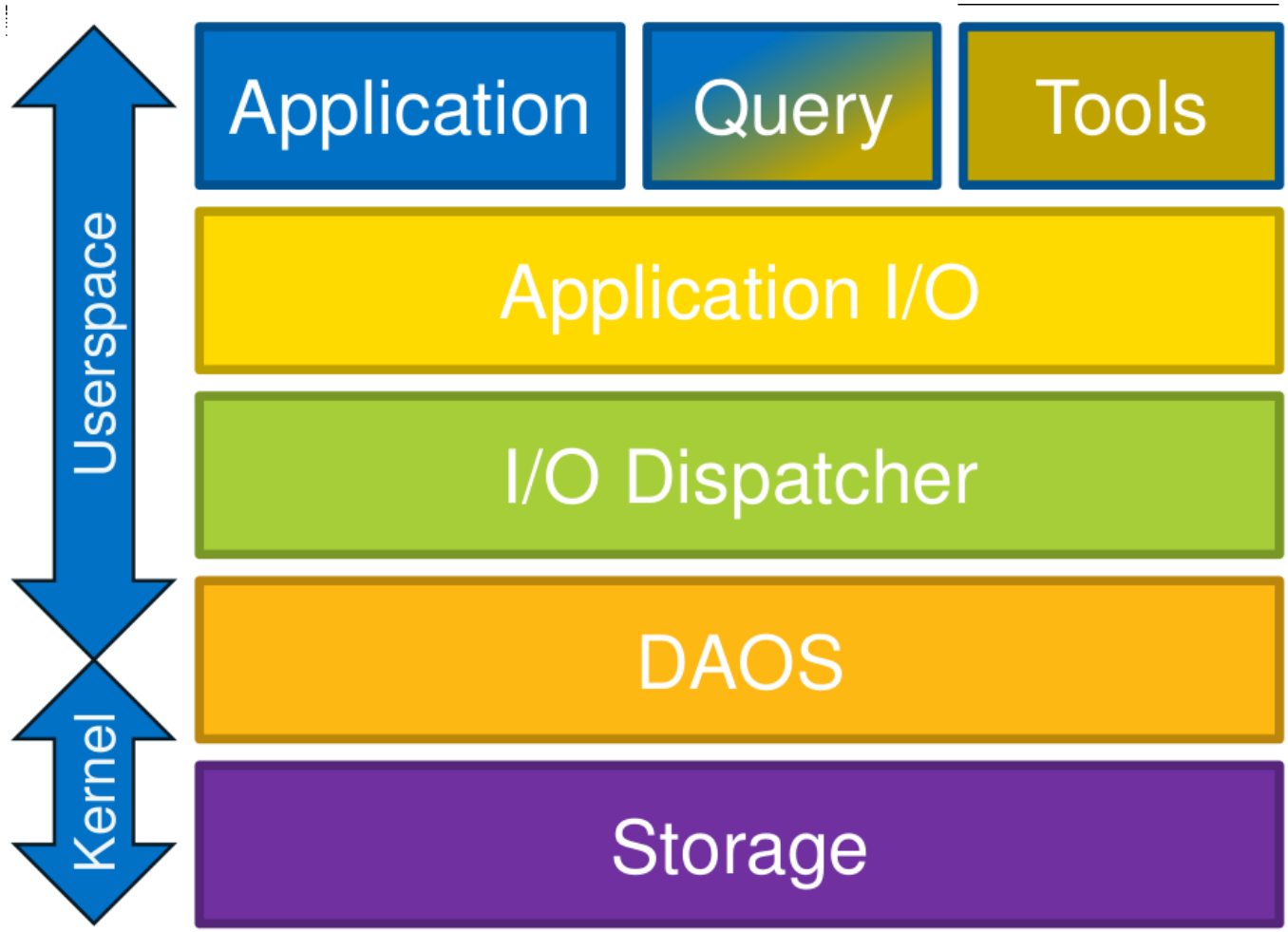
Figure courtesy Intel.

SWARM for DAGs: hardware implementation



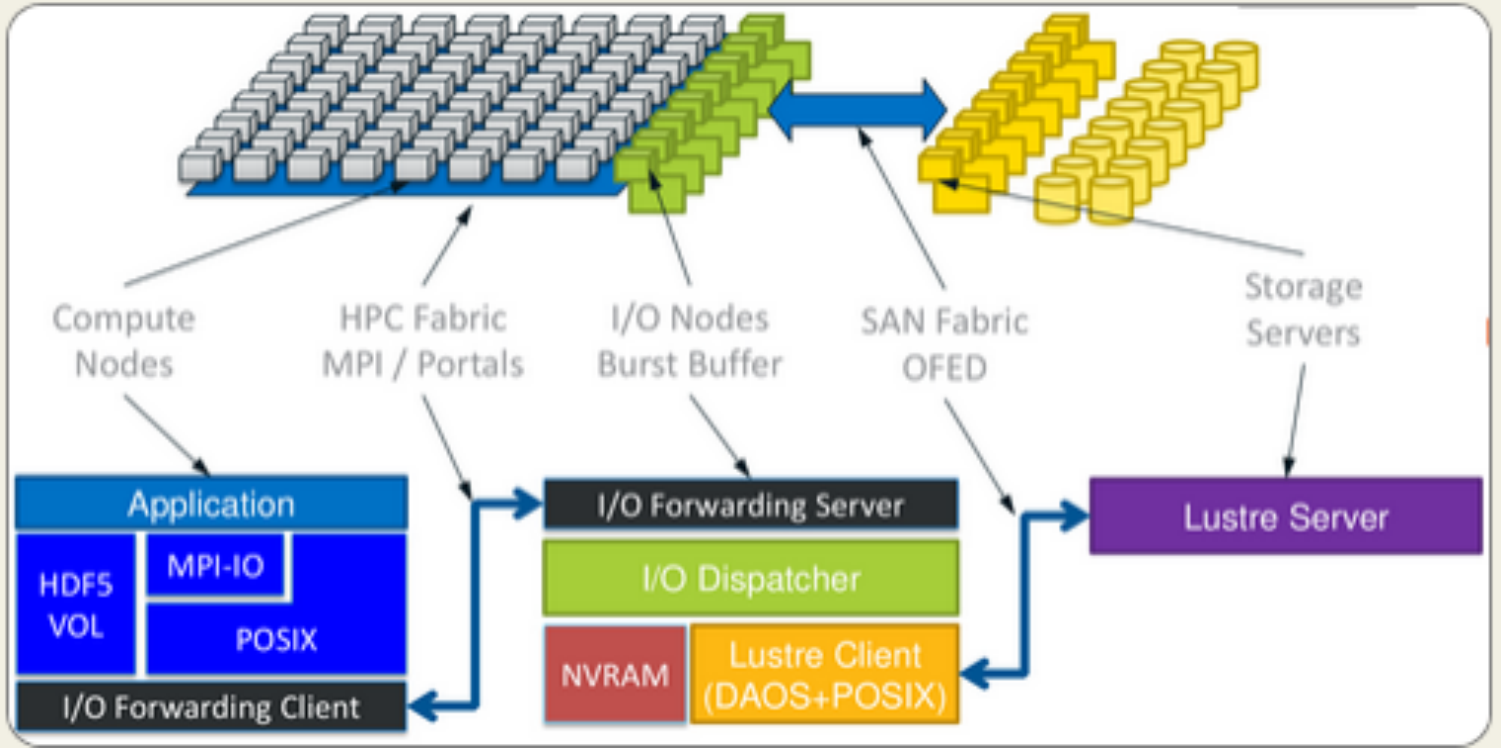
Jeffrey et al, *IEEE Micro* 2016.

NVRAM will blur distinction between memory and filesystem



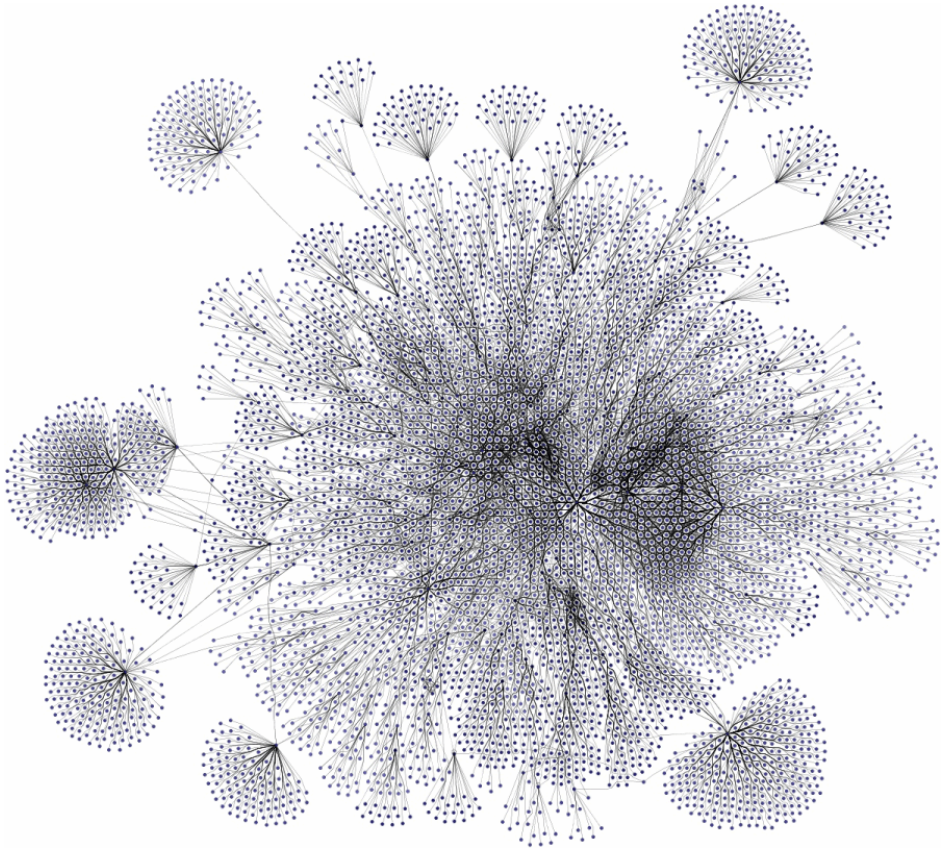
Hemsoth, 2014: **http://goo.gl/3ZeOXt**

NVRAM will blur distinction between memory and filesystem



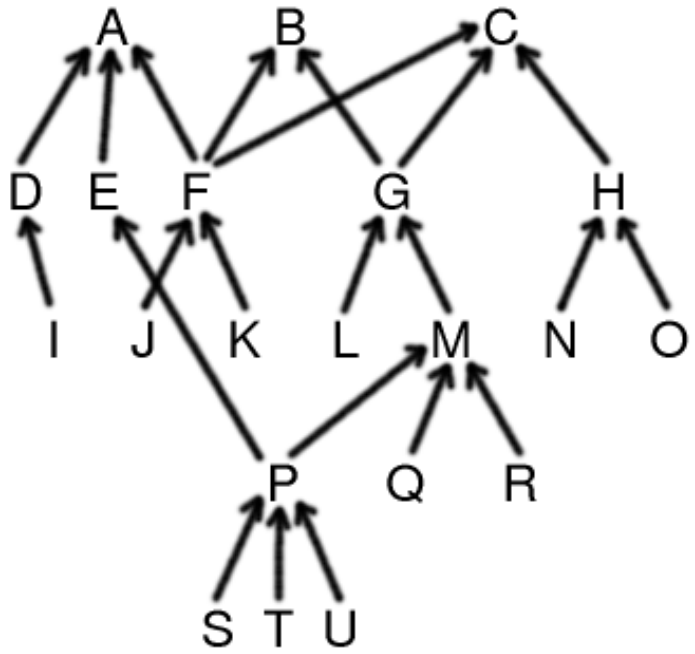
Hemsoth, 2014: **http://goo.gl/3ZeOXt**

Work avoidance



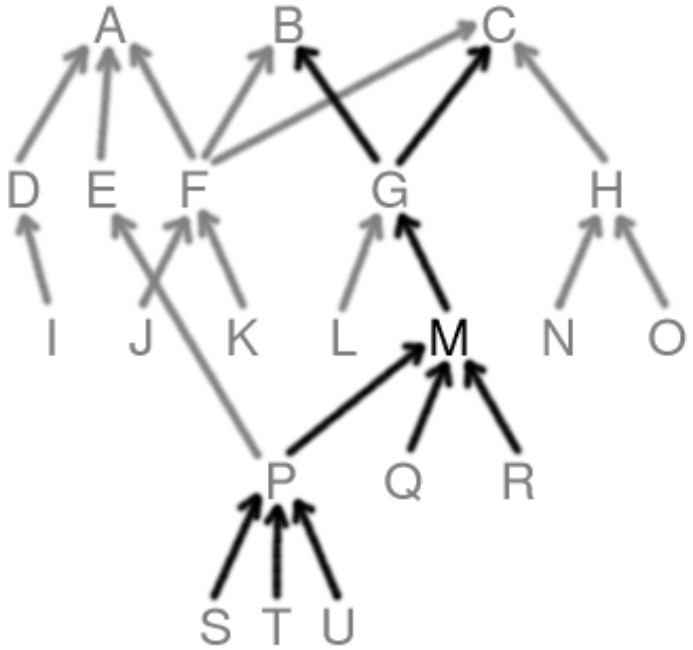
- Work avoidance: find minimal path to complete output
  - **`make`**: traverse tree backwards; state is the filesystem state.
  - **`cylc/chaco`**: traverse tree forwards; each task formulated as a no-op if outputs exist; **`fred`** contains state including tasks in flight.

Work avoidance



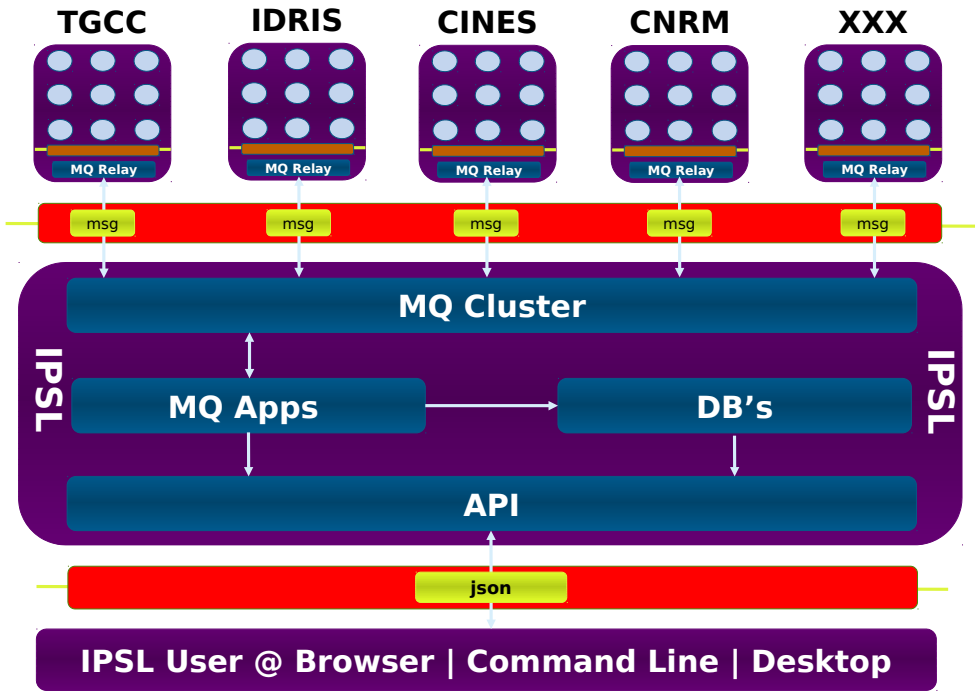
- Work avoidance: find minimal path to complete output
  - **`make`**: traverse tree backwards; state is the filesystem state.
  - **`cylc/chaco`**: traverse tree forwards; each task formulated as a no-op if outputs exist; **`fred`** contains state including tasks in flight.

Work avoidance



- Work avoidance: find minimal path to complete output
  - **`make`**: traverse tree backwards; state is the filesystem state.
  - **`cylc/chaco`**: traverse tree forwards; each task formulated as a no-op if outputs exist; **`fred`** contains state including tasks in flight.

Use of cross-network message queues



- IPSL have tested handling $\mathcal{O}(10^5)$ enqueues/dequeues per day.
- Google reports Rabbit service of $\mathcal{O}(10^6)$ *per second*! (more than all SMS/WhatsApp/etc) **https://goo.gl/GBlAAz**
- AMQP: active messages containing instructions as well as data.

Figure courtesy Sébastien Denvil, IPSL.

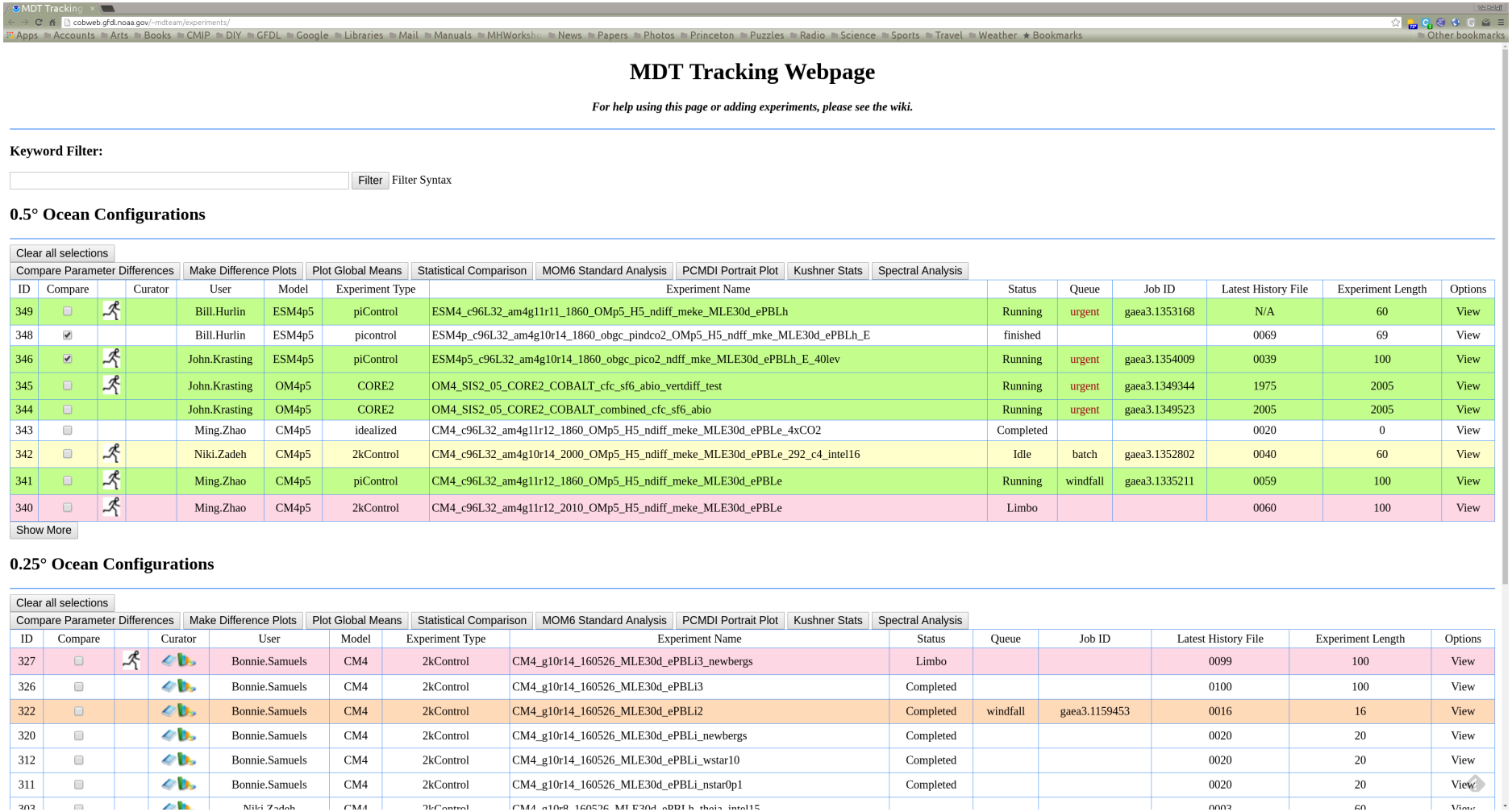
Outline

Development and production workflow

Model developers have different workflow priorities and requirements.

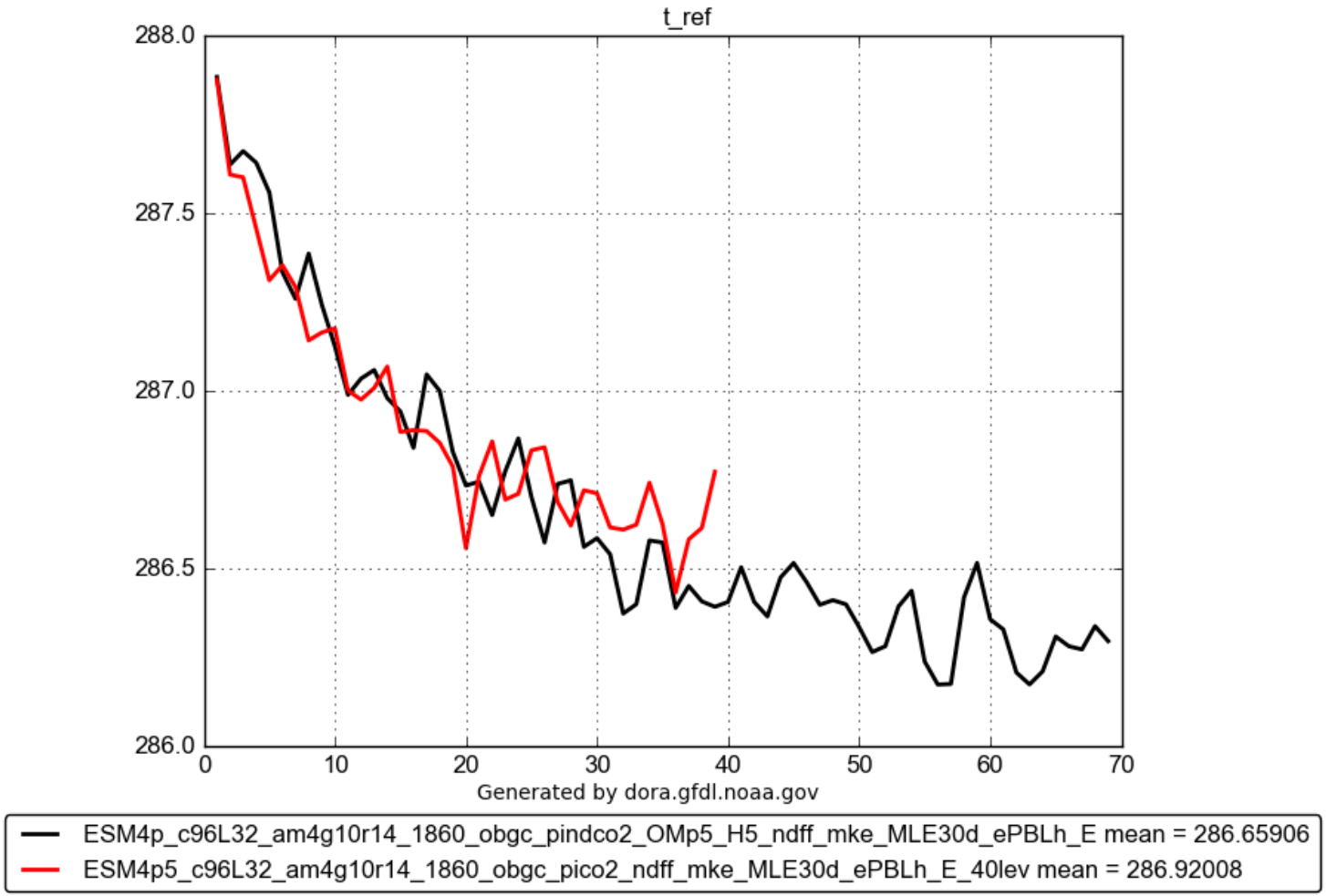
- Production workflow benefits from coherence and similarity across runs.
- Development workflow requires extremely fine-grained access to code, namelists, scripts. A lot of rules broken:
  - Favored IDE/UI is called **`vi`**!
  - source code edits in user directories
  - input file modifications on the fly
- Analysis workflow requires random access to local disk: inspiration-driven rather than industrial strength
- Still benefit from regression testing harness: multiple compilers, platforms
- Emulators? e.g SoftFloat **`http://www.jhauser.us/arithmetic/SoftFloat.html`**
- Provenance and metadata requirements relaxed for development workflow.

Statistical comparison across model versions



Live monitoring of model runs. From GFDL MDT Tracking Page...

Statistical comparison across model versions



Are two runs the same or different? What difference in inputs is responsible for the disrepancy? From GFDL MDT Tracking Page...

Multi-model ensembles for climate projection

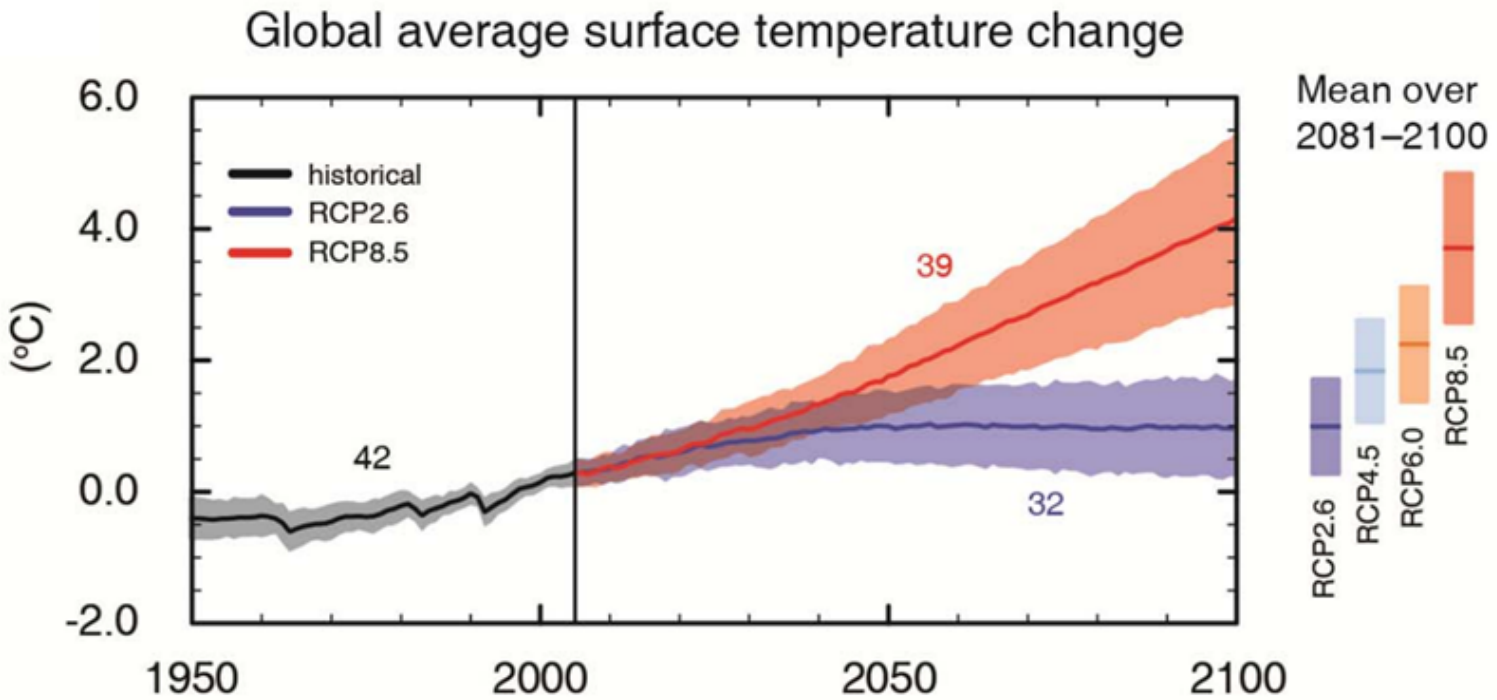
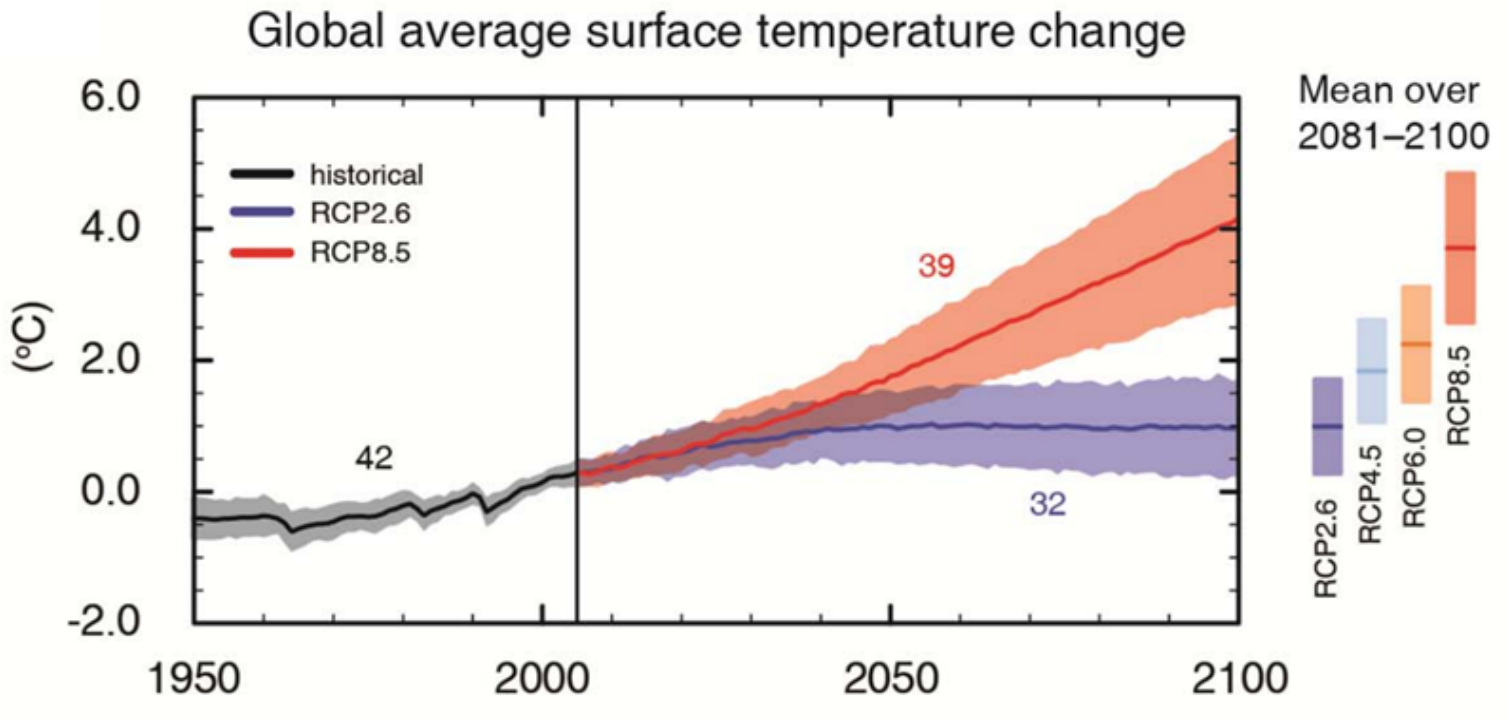


Figure SPM.7 from the IPCC AR5 Report. Can be interpreted as the most general and rigorous test of scientific reproducibility.

Multi-model ensembles for climate projection



- Critically depends on software, metadata, and data standards: the Earth System Grid Federation (**http://esgf.org**): a 3 PB federated archive.
- Workflows for replication, versioning, subsetting, QC, citation.

Outline

Summary

- Community is struggling to move from synchronous to asynchronous data flow on the coming hardware platforms.
- Hardware is blurring the line between cache and memory, memory and storage (deep hierarchy).
- Energy to solution as a benchmark across the entire workflow.
- Expressing entire workflow as graphs (DAGs).
  - Maximize parallelism across the entire graph
  - Minimize graph traversal during fault recovery
- Accommodating different needs for development and production workflows: relax provenance and metadata requirements during development.
- Irreproducible computing: include statistical consistency testing into workflow.

Summary

- Community is struggling to move from synchronous to asynchronous data flow on the coming hardware platforms.
- Hardware is blurring the line between cache and memory, memory and storage (deep hierarchy).
- Energy to solution as a benchmark across the entire workflow.
- Expressing entire workflow as graphs (DAGs).
  - Maximize parallelism across the entire graph
  - Minimize graph traversal during fault recovery
- Accommodating different needs for development and production workflows: relax provenance and metadata requirements during development.
- Irreproducible computing: include statistical consistency testing into workflow.
- CMIP8 is going to be awesome!

Thank you

Thank you.