

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

1 Hardware Directions

- GPUs, MICs, ARM
- Inexact computing
- Energy cost of algorithms and data movement

2 A Graph Approach

- Directed Acyclic Graphs
- Convergence of computation and data
- Fault tolerance across the workflow

3 Metadata and provenance

- Development and production workflow
- Statistical and scientific reproducibility

4 Summary

Outline

1 Hardware Directions

- GPUs, MICs, ARM
- Inexact computing
- Energy cost of algorithms and data movement

2 A Graph Approach

- Directed Acyclic Graphs
- Convergence of computation and data
- Fault tolerance across the workflow

3 Metadata and provenance

- Development and production workflow
- Statistical and scientific reproducibility

4 Summary

Power-8 with NVLink

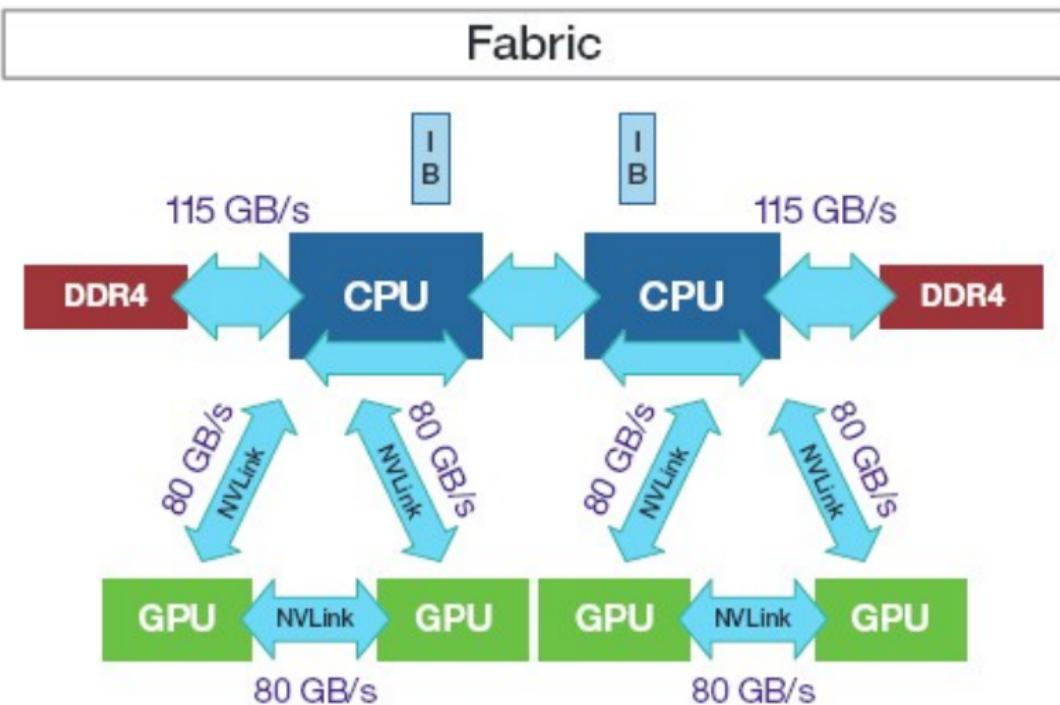
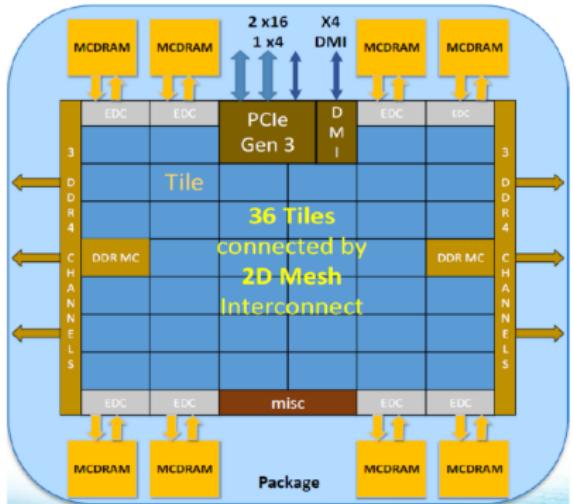


Figure courtesy IBM.

KNL Overview

KNL Overview



TILE



Chip: 36 Tiles interconnected by 2D Mesh
Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW
DDR4: 6 channels @ 2400 up to 384 GB

IO: 36 lanes PCIe* Gen3. 4 lanes of DMI for chipset

Node: 1-Socket only

Fabric: Intel® Omni-Path Architecture on-package (not shown)

MCDRAM
~5X Higher BW
than DDR

Vector Peak Perf: 3+TF DP and 6+TF SP Flops

Scalar Perf: ~3x over Knights Corner

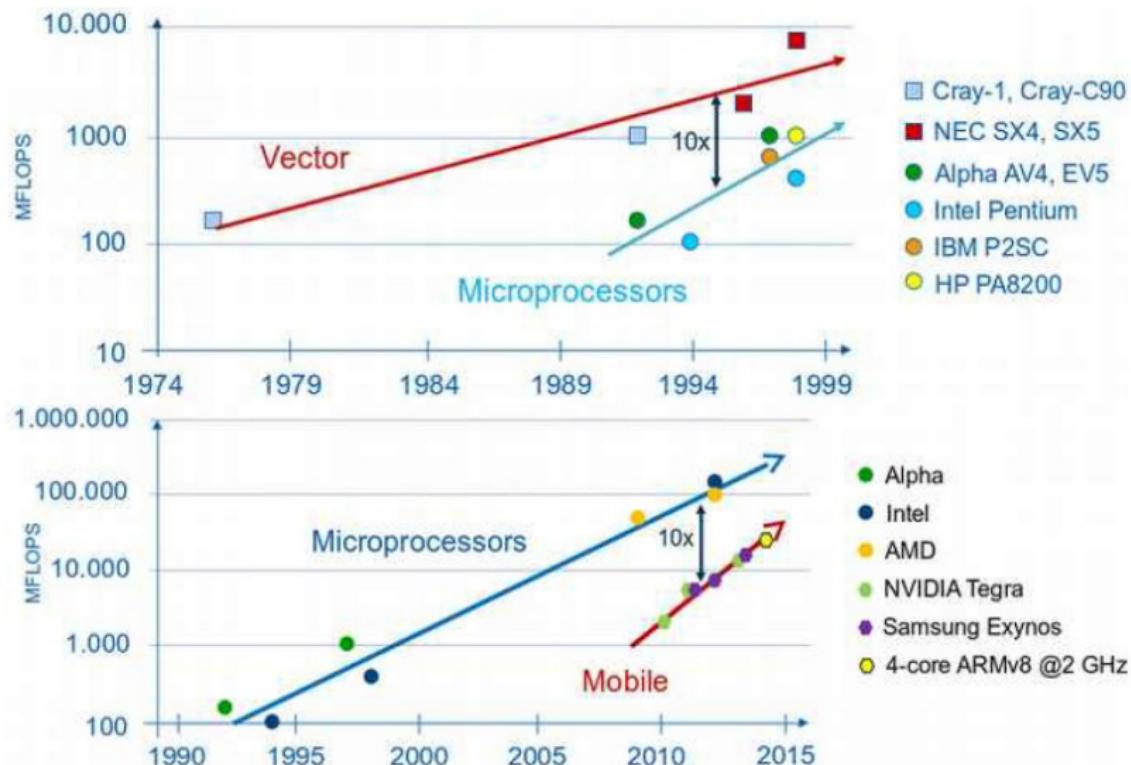
Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

* Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). 2Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. *Other names and brands may be claimed as the property of others.



Figure courtesy Intel.

The inexorable triumph of commodity computing



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

Irreproducible Computing, Inexact Hardware

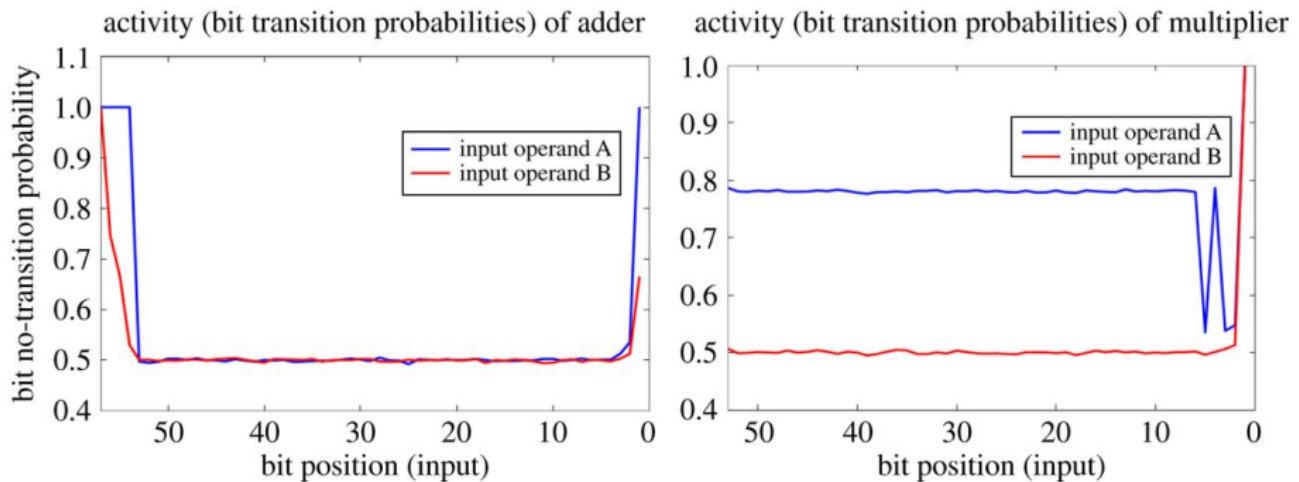


Figure 1 from Düben et al, *Phil. Trans. A*, 2016. Which bits can we allow to be “inexactly” flipped? Lorenz 96 as canonical test case of non-linearity and chaos.

Irreproducible Computing, Inexact Hardware

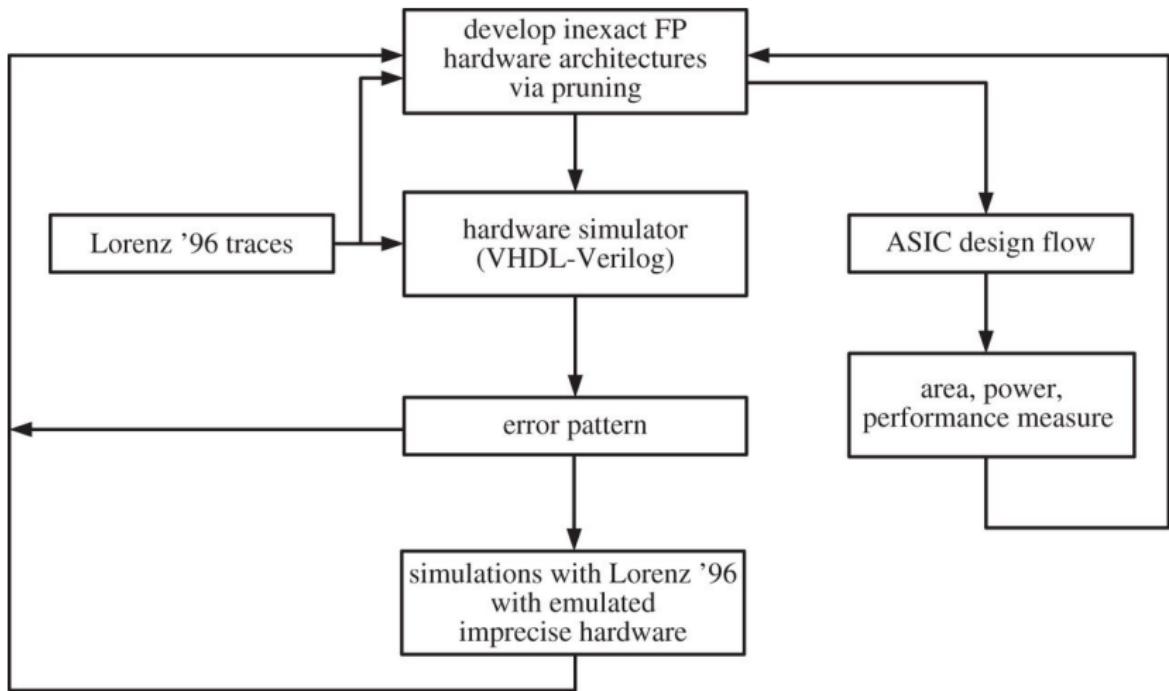


Figure 2 from Düben et al, *Phil. Trans. A*, 2016.

COSMO: energy to solution

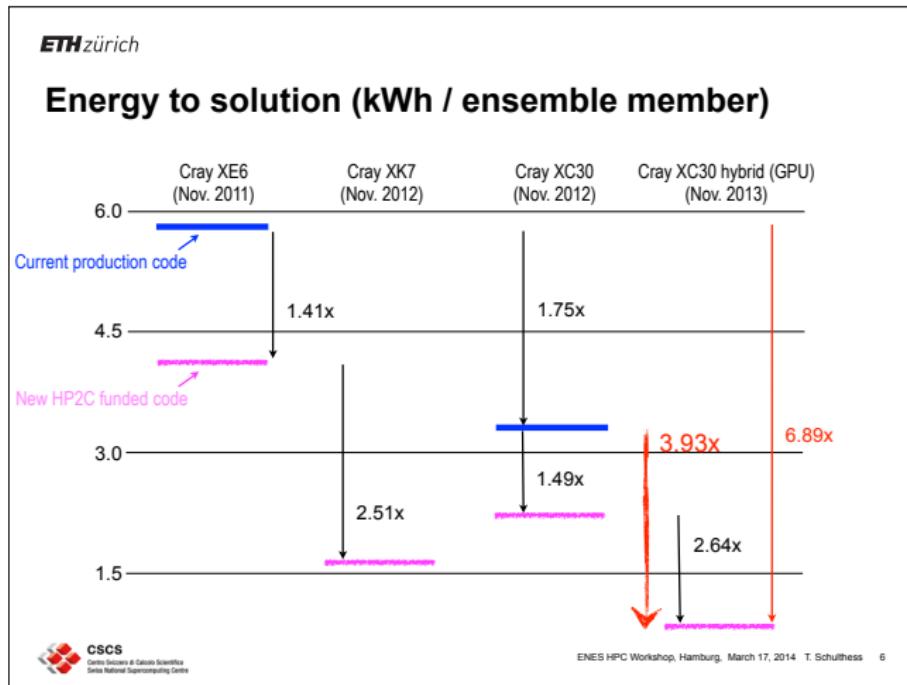


Figure courtesy Thomas Schulthess, CSCS.

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

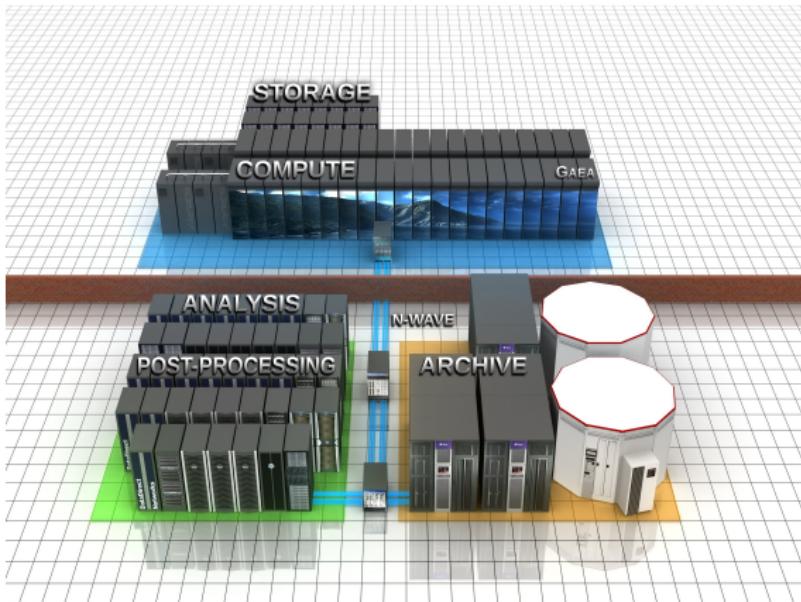
- Comparative measures of capability (SYPD), capacity (CHSY), and energy cost (JPSY) per “unit of science”.
- Can you have codes that are “slower but greener”? Algorithms that are “less accurate but more eco-friendly”?
- From Balaji et al (2016), in review at GMDD.

<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 (Schultess).
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 ~ 10 pJ to main memory; ~ 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

1 Hardware Directions

- GPUs, MICs, ARM
- Inexact computing
- Energy cost of algorithms and data movement

2 A Graph Approach

- Directed Acyclic Graphs
- Convergence of computation and data
- Fault tolerance across the workflow

3 Metadata and provenance

- Development and production workflow
- Statistical and scientific reproducibility

4 Summary

Examples of DAG parallelism

ECMWF Seminar 2013

DAG example: Cholesky Inversion

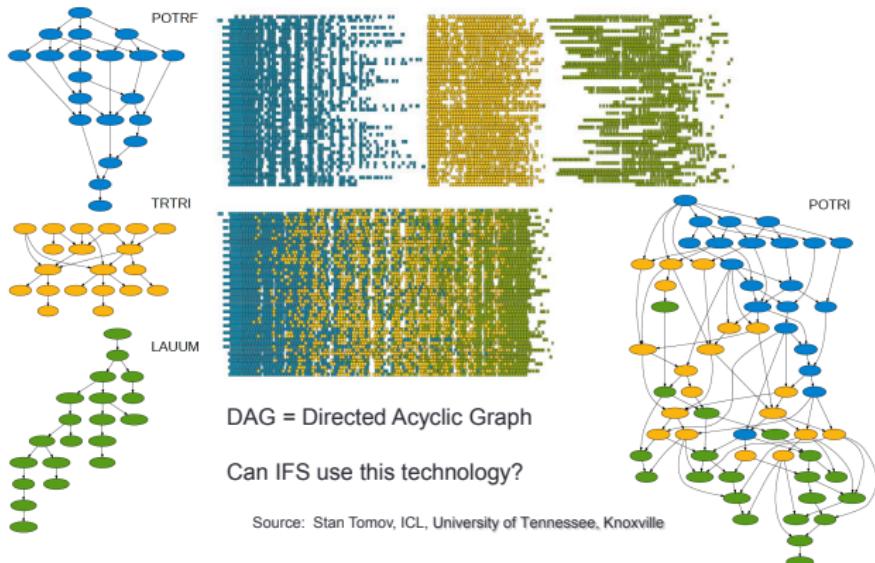
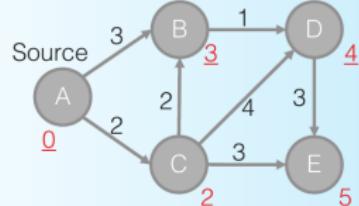
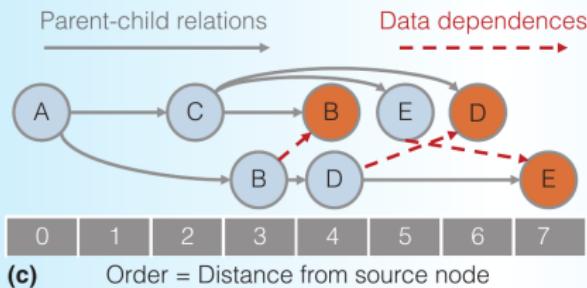


Figure courtesy George Mozdzynski, ECMWF.

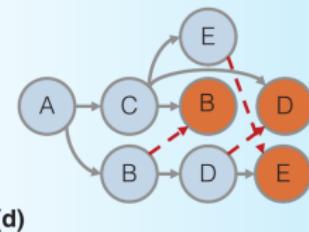
SWARM for DAGs

```
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)

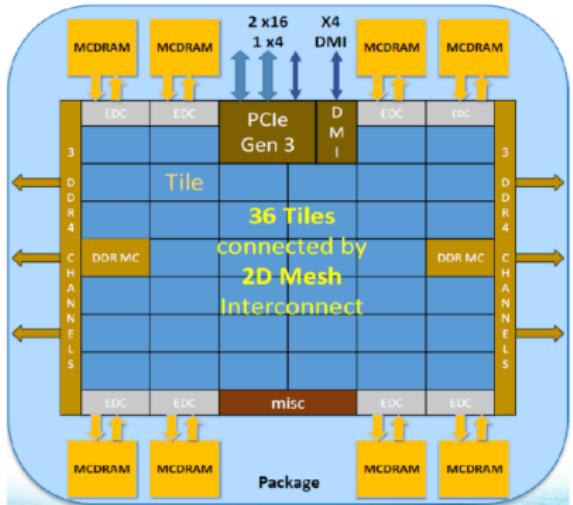


(d)

Jeffrey et al, IEEE Micro 2016.

KNL Overview

KNL Overview



TILE



Chip: 36 Tiles interconnected by 2D Mesh
Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW
DDR4: 6 channels @ 2400 up to 384 GB

IO: 36 lanes PCIe* Gen3. 4 lanes of DMI for chipset

Node: 1-Socket only

Fabric: Intel® Omni-Path Architecture on-package (not shown)

Vector Peak Perf: 3+TF DP and 6+TF SP Flops

Scalar Perf: ~3x over Knights Corner

Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

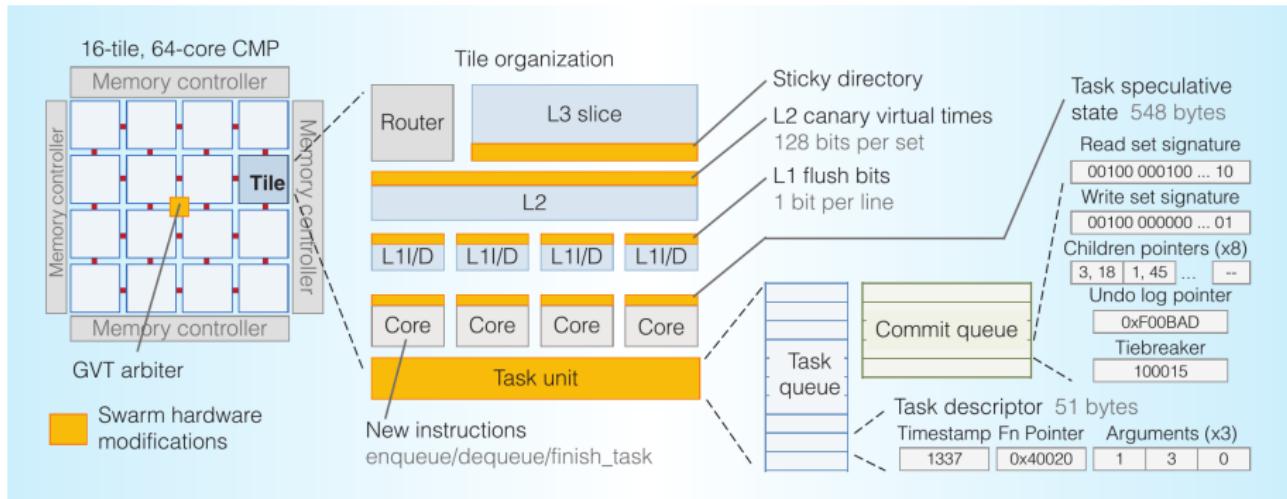
MCDRAM
~5X Higher BW
than DDR

* Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). 2Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. *Other names and brands may be claimed as the property of others.



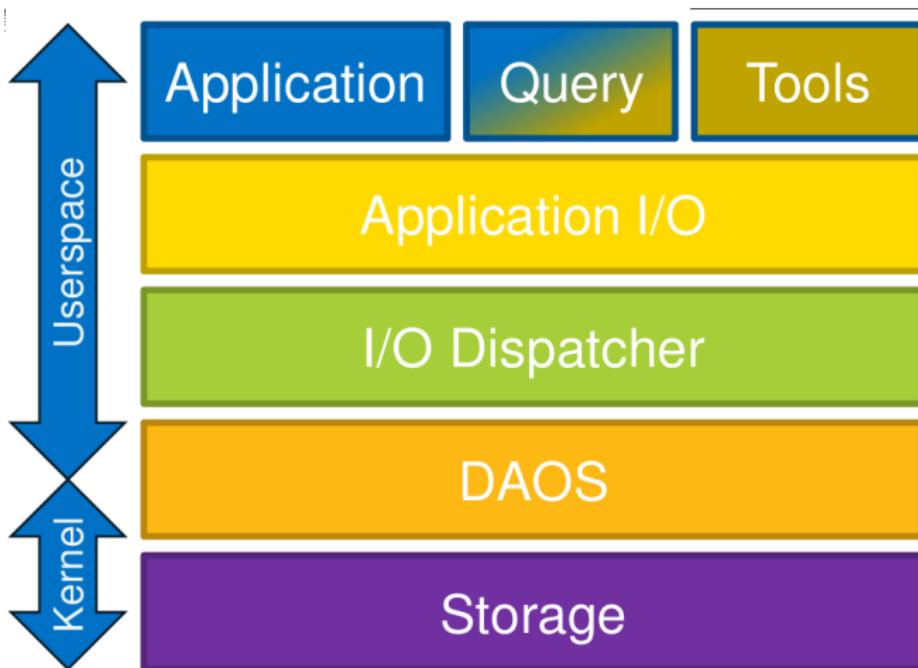
Figure courtesy Intel.

SWARM for DAGs: hardware implementation



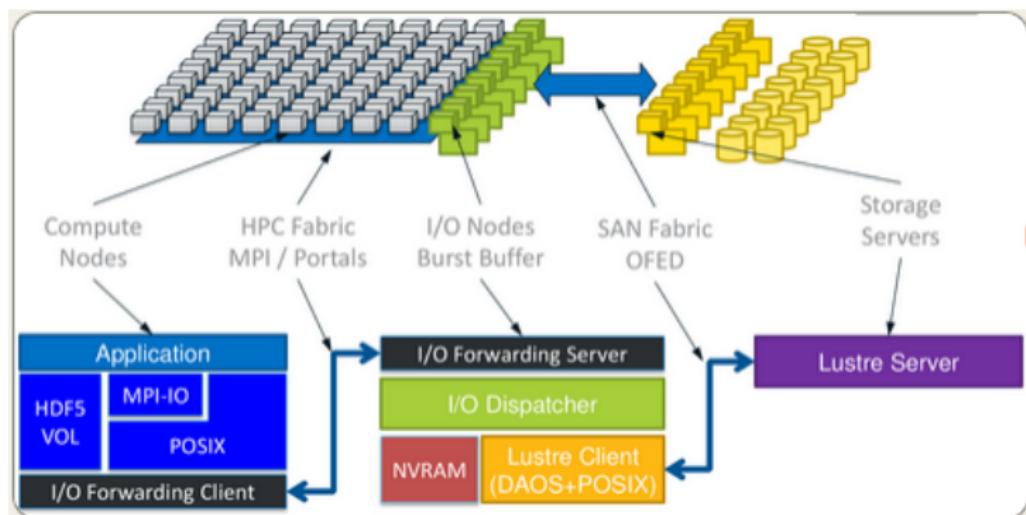
Jeffrey et al, *IEEE Micro* 2016.

NVRAM will blur distinction between memory and filesystem



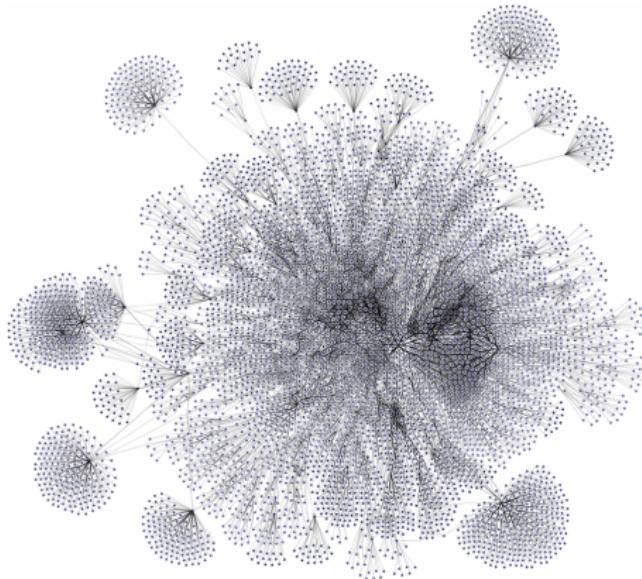
Hemsoth, 2014: <http://goo.gl/3ZeOxt>

NVRAM will blur distinction between memory and filesystem



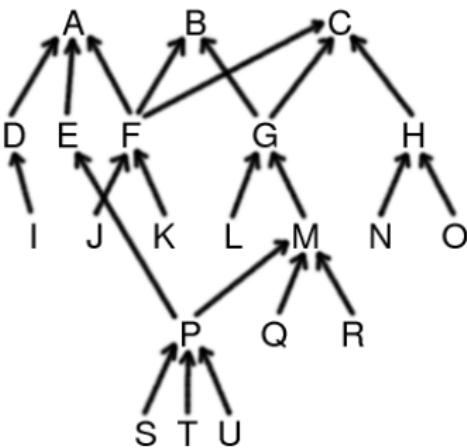
Hemsoth, 2014: <http://goo.gl/3ZeOxt>

Work avoidance



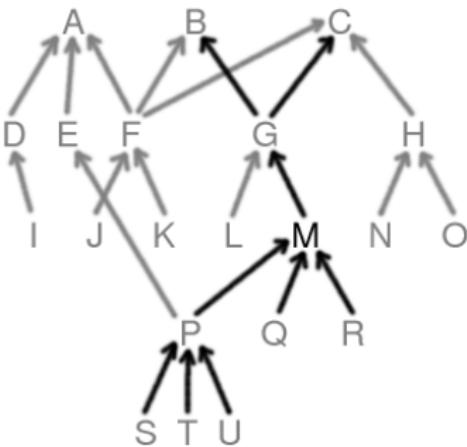
- Work avoidance: find minimal path to complete output
 - **make**: traverse tree backwards; state is the filesystem state.
 - **cylc/chaco**: traverse tree forwards; each task formulated as a **no-op** if outputs exist; **fred** contains state **including tasks in flight**.

Work avoidance



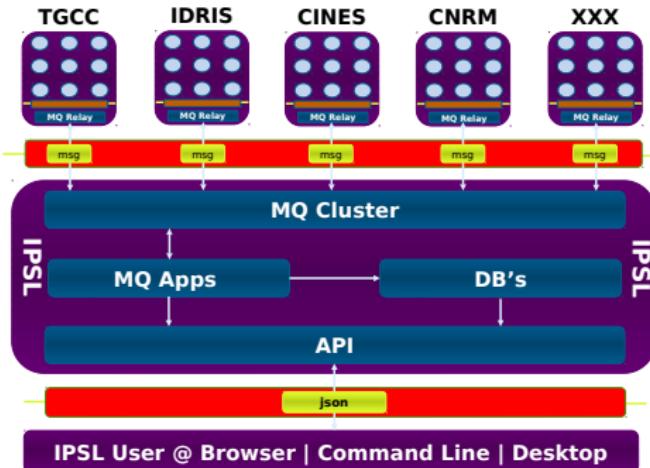
- Work avoidance: find minimal path to complete output
 - `make`: traverse tree backwards; state is the filesystem state.
 - `cylc/chaco`: traverse tree forwards; each task formulated as a `no-op` if outputs exist; `fred` contains state **including tasks in flight**.

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/GB1AAz>
- AMQP: **active messages** containing **instructions** as well as data.

Figure courtesy Sébastien Denvil, IPSL.

Outline

1 Hardware Directions

- GPUs, MICs, ARM
- Inexact computing
- Energy cost of algorithms and data movement

2 A Graph Approach

- Directed Acyclic Graphs
- Convergence of computation and data
- Fault tolerance across the workflow

3 Metadata and provenance

- Development and production workflow
- Statistical and scientific reproducibility

4 Summary

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

MOT Track Webpage

MDT Tracking Webpage

For help using this page or adding experiments, please see the wiki.

Keyword Filter: Filter Syntax

0.5° Ocean Configurations

Clear all selections

ID	Compare	Curator	User	Model	Experiment Type	Experiment Name	Status	Queue	Job ID	Latest History File	Experiment Length	Options
349	Bill.Hurlin	ESM4p5	pControl	ESM4_c96L32_am4g11r11_1860_Omp5_H5_ndiff_meko_MLE30d_ePBLh	Running	urgent	gaoa3.135316B	N/A	60	View		
348	Bill.Hurlin	ESM4p5	pControl	ESM4p_c96L32_am4g10r14_1860_omp5_H5_ndiff_meko_MLE30d_ePBLh_E	finished			0099	69	View		
346	John.Krusting	ESM4p5	pControl	ESM4p5_c96L32_am4g10r14_1860_omp5_H5_ndiff_meko_MLE30d_ePBLh_E_40ev	Running	urgent	gaoa3.1354009	0099	100	View		
345	John.Krusting	OM4p5	CORE2	OM4_SIS2_05_CORE2_COBALT_cfc_sfc_abo_vendiff_test	Running	urgent	gaoa3.1349344	1975	2005	View		
344	John.Krusting	OM4p5	CORE2	OM4_SIS2_05_CORE2_COBALT_cfc_sfc_abi	Running	urgent	gaoa3.1349523	2005	2005	View		
343	Ming.Zhao	CM4p5	idealized	CM4_c96L32_am4g11r12_1860_Omp5_H5_ndiff_meko_MLE30d_ePBLe_AxC02	Completed			0020	0	View		
342	Niki.Zadeh	CM4p5	2kControl	CM4_c96L32_am4g10r14_2000_Omp5_H5_ndiff_meko_MLE30d_ePBL_e_292_c4_int16	Idle	batch	gaoa3.1352802	0040	60	View		
341	Ming.Zhao	CM4p5	pControl	CM4_c96L32_am4g11r12_1860_Omp5_H5_ndiff_meko_MLE30d_ePBLe	Running	windfall	gaoa3.1335211	0059	100	View		
340	Ming.Zhao	CM4p5	2kControl	CM4_c96L32_am4g11r12_2010_Omp5_H5_ndiff_meko_MLE30d_ePBLe	Limbo			0060	100	View		

Show More

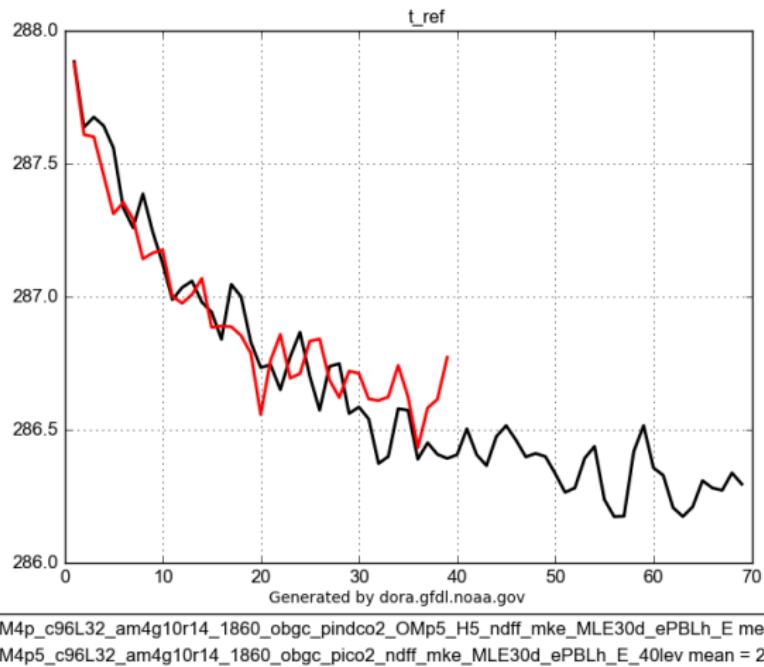
0.25° Ocean Configurations

Clear all selections

ID	Compare	Curator	User	Model	Experiment Type	Experiment Name	Status	Queue	Job ID	Latest History File	Experiment Length	Options
327	Bonnie.Samuels	CM4	2kControl	CM4_g10r14_160526_MLE30d_ePBL3_newbergs	Limbo			0099	100	View		
326	Bonnie.Samuels	CM4	2kControl	CM4_g10r14_160526_MLE30d_ePBL3	Completed			0100	100	View		
322	Bonnie.Samuels	CM4	2kControl	CM4_g10r14_160526_MLE30d_ePBL2	Completed	windfall	gaoa3.1159453	0016	16	View		
320	Bonnie.Samuels	CM4	2kControl	CM4_g10r14_160526_MLE30d_ePBL3_newbergs	Completed			0020	20	View		
312	Bonnie.Samuels	CM4	2kControl	CM4_g10r14_160526_MLE30d_ePBL1_wstar10	Completed			0020	20	View		
311	Bonnie.Samuels	CM4	2kControl	CM4_g10r14_160526_MLE30d_ePBL1_star10p1	Completed			0020	20	View		
309	All	All	All	All	All	All			0000	000	More	

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 discrepancy? From GFDL MDT Tracking Page...

Multi-model ensembles for climate projection

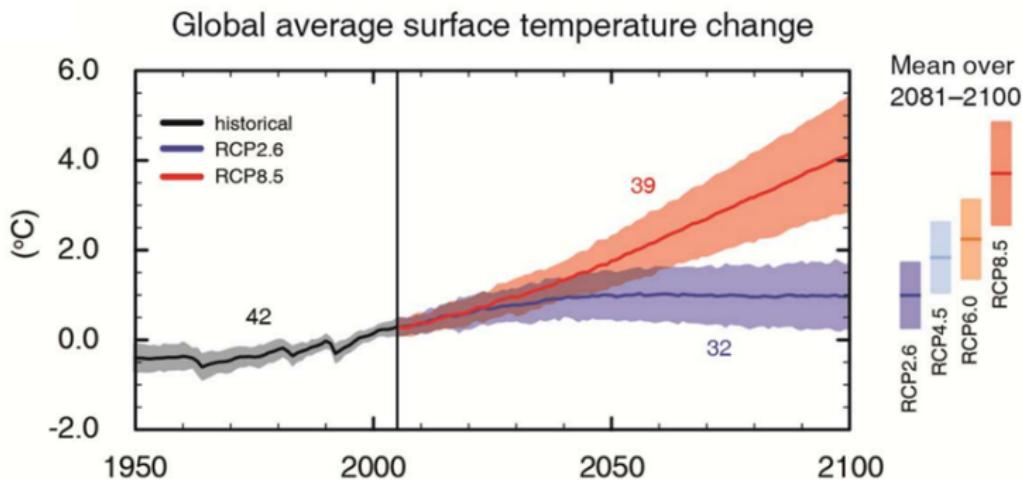
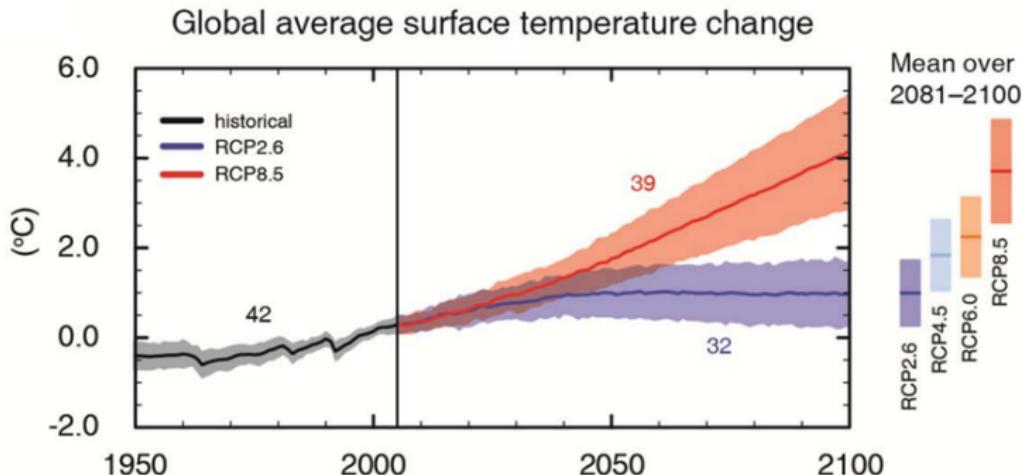


Figure SPM.7 from the IPCC AR5 Report. Can be interpreted as the most general and rigorous test of **scientific reproducibility**.

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

1

Hardware Directions

- GPUs, MICs, ARM
- Inexact computing
- Energy cost of algorithms and data movement

2

A Graph Approach

- Directed Acyclic Graphs
- Convergence of computation and data
- Fault tolerance across the workflow

3

Metadata and provenance

- Development and production workflow
- Statistical and scientific reproducibility

4

Summary

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.