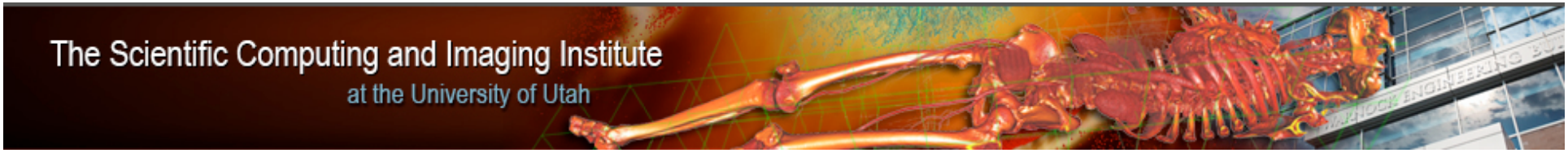


On the Future of High Performance Computing: How to Think for Peta and Exascale Computing

Jack Dongarra

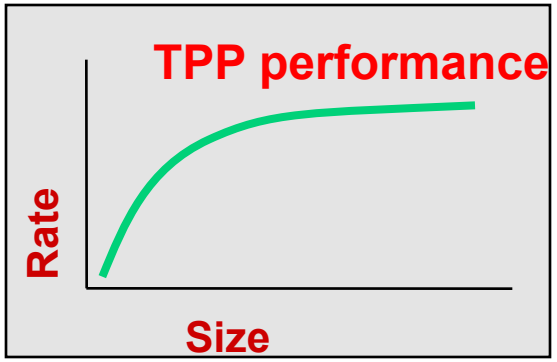
University of Tennessee
Oak Ridge National Laboratory
University of Manchester

Top500 List of Supercomputers

H. Meuer, H. Simon, E. Strohmaier, & JD

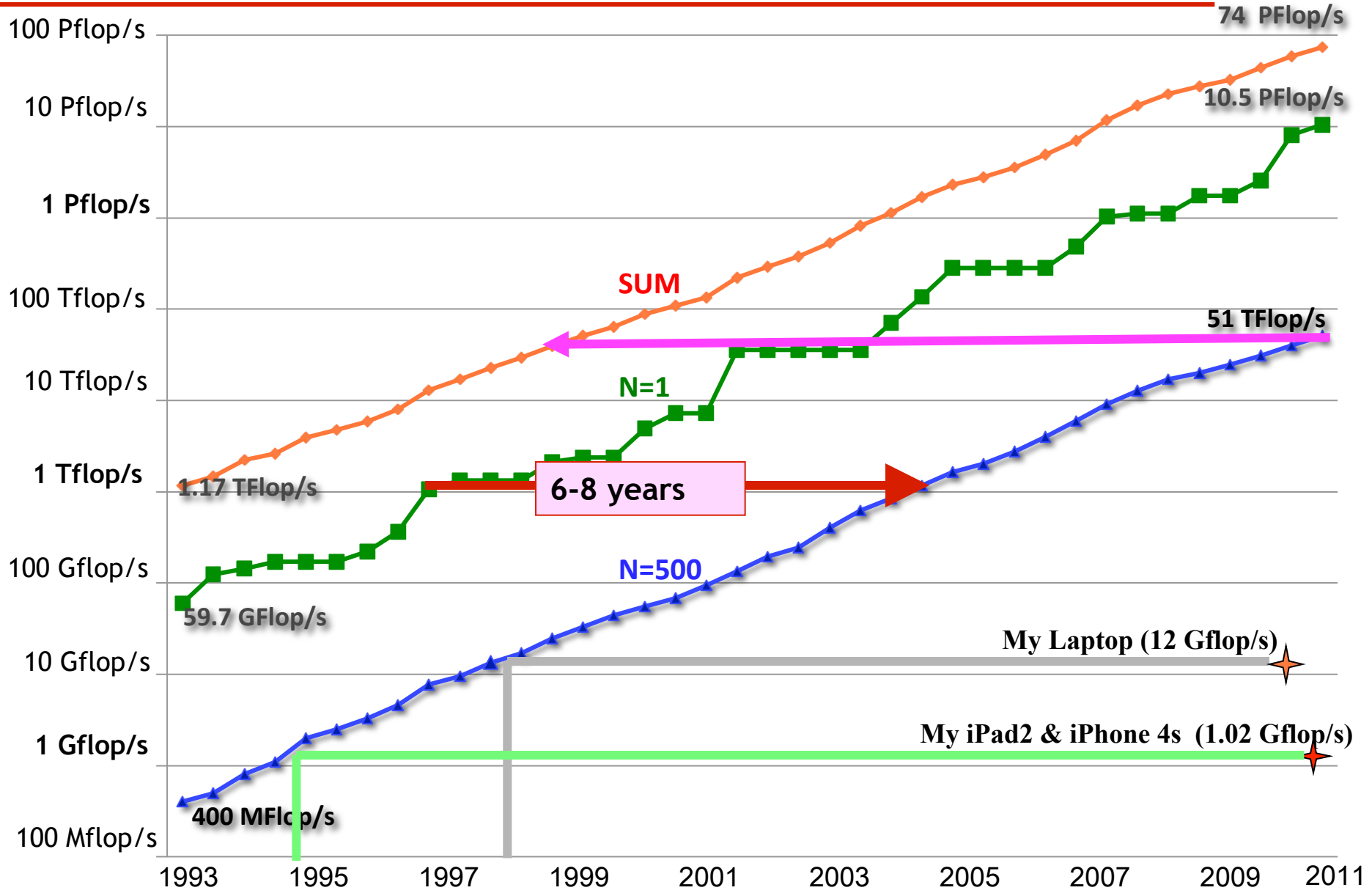
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

$Ax=b$, *dense problem*

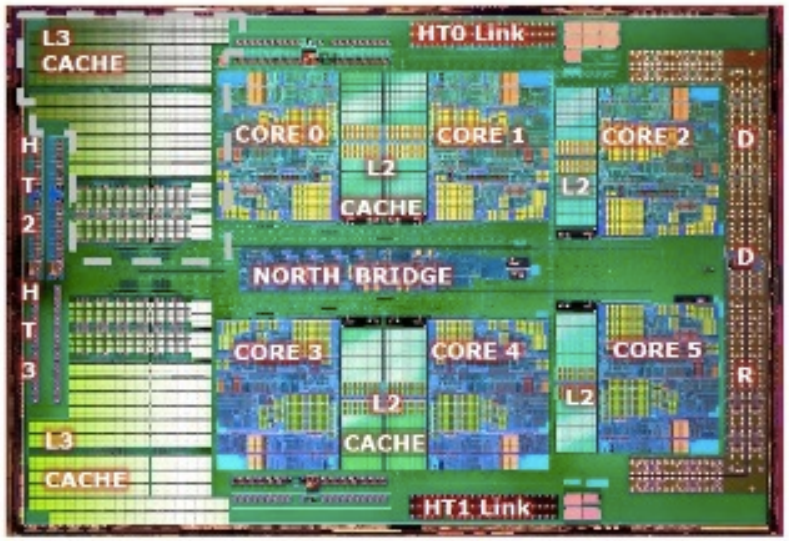


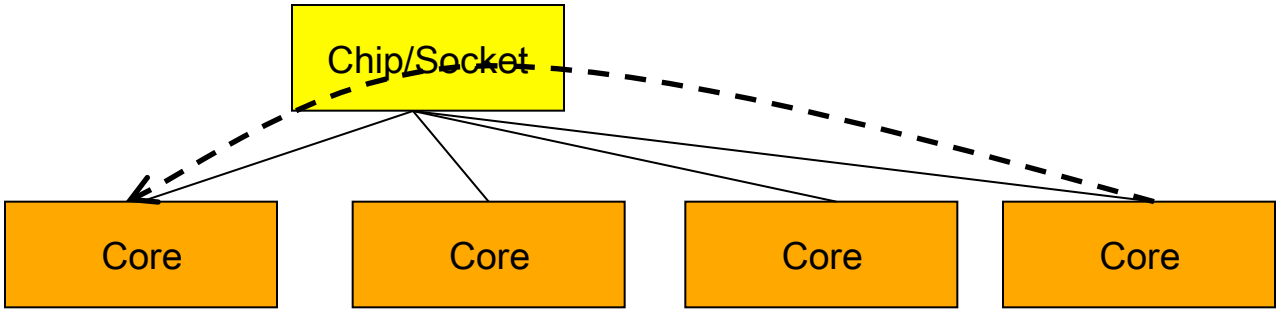
- Updated twice a year
  SC‘xy in the States in November
  Meeting in Germany in June
- All data available from **www.top500.org**

Performance Development

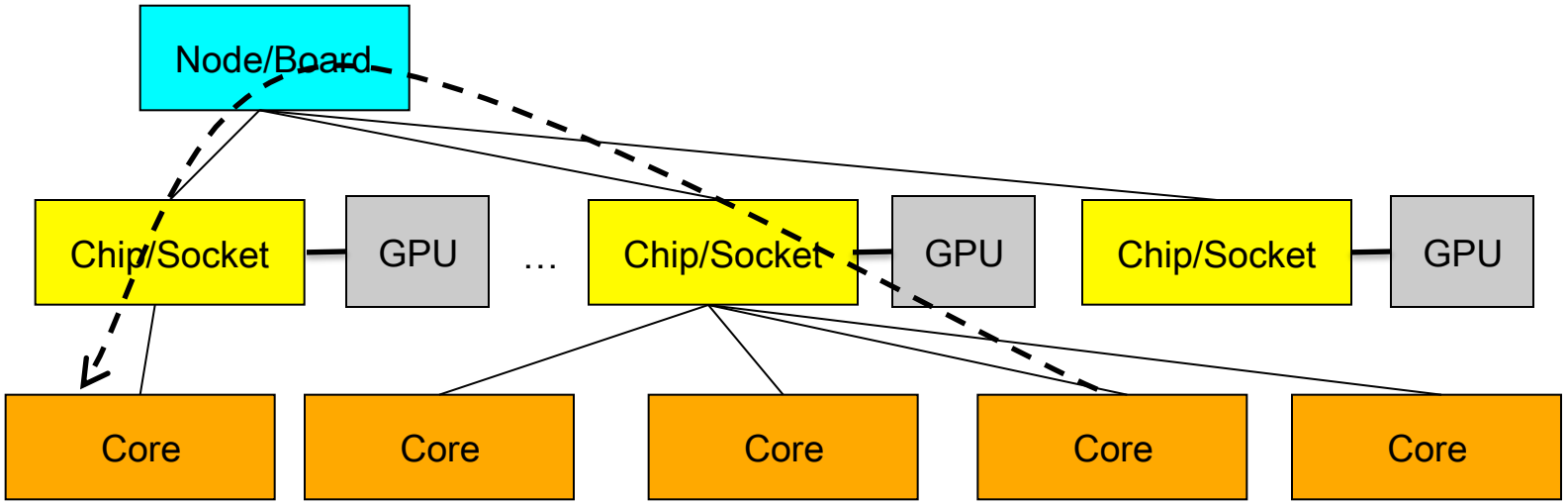


Example of typical parallel machine



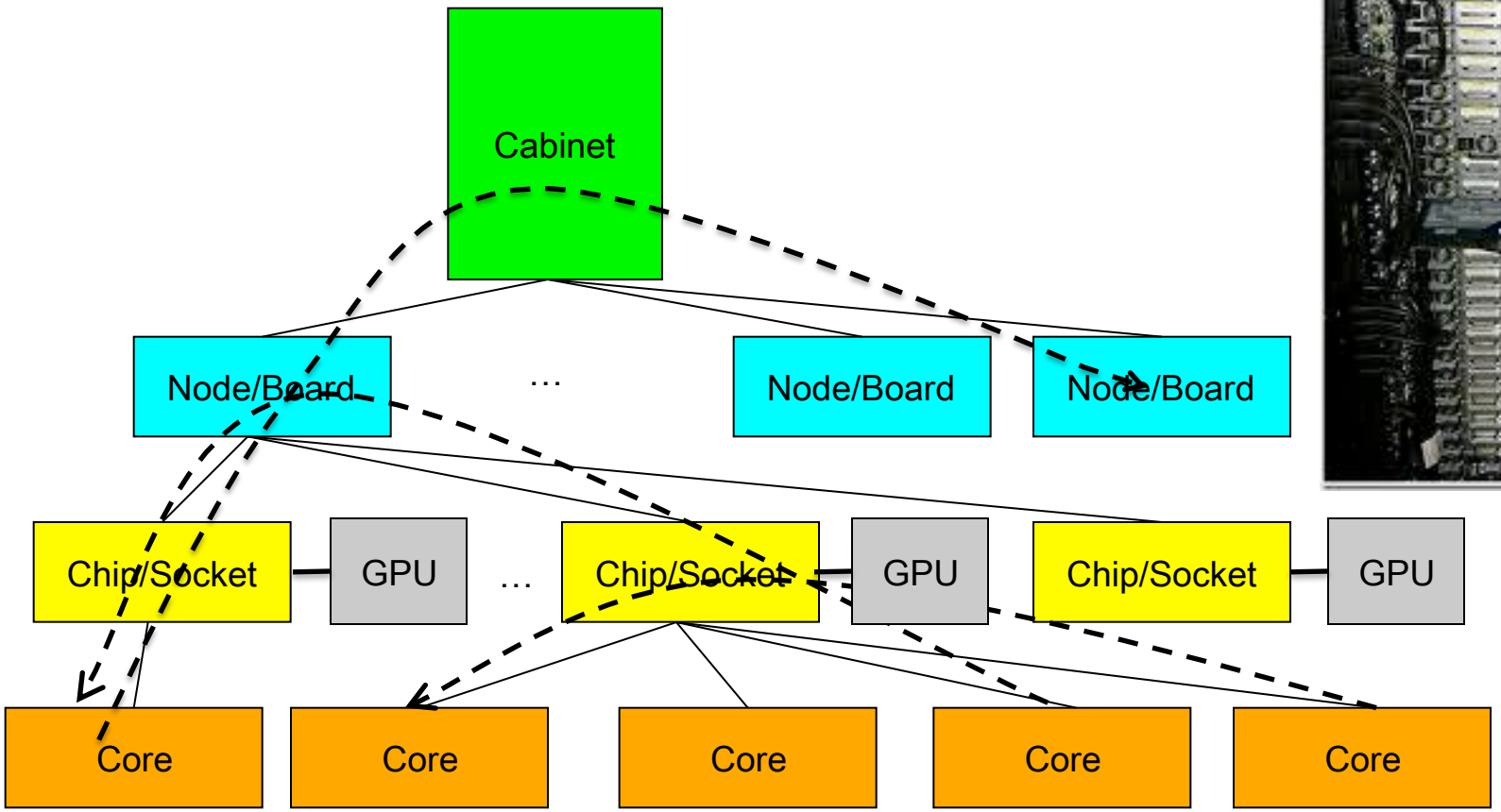


Example of typical parallel machine

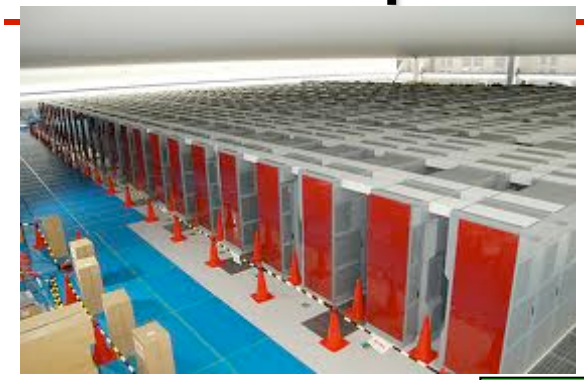


Example of typical parallel machine

Shared memory programming between processes on a board and
a combination of shared memory and distributed memory programming
between nodes and cabinets



Example of typical parallel machine



Combination of shared memory and distributed memory programming

November 2011: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak
1	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx + custom	Japan	705,024	10.5	93
2	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel + Nvidia GPU + custom	China	186,368	2.57	55
3	DOE / OS Oak Ridge Nat Lab	Jaguar, Cray AMD + custom	USA	224,162	1.76	75
4	Nat. Supercomputer Center in Shenzhen	Nebulea, Dawning Intel + Nvidia GPU + IB	China	120,640	1.27	43
5	GSIC Center, Tokyo Institute of Technology	Tusbame 2.0, HP Intel + Nvidia GPU + IB	Japan	73,278	1.19	52
6	DOE / NNSA LANL & SNL	Cielo, Cray AMD + custom	USA	142,272	1.11	81
7	NASA Ames Research Center/NAS	Plelades SGI Altix ICE 8200EX/8400EX + IB	USA	111,104	1.09	83
8	DOE / OS Lawrence Berkeley Nat Lab	Hopper, Cray AMD + custom	USA	153,408	1.054	82
9	Commissariat a l'Energie Atomique (CEA)	Tera-10, Bull Intel + IB	France	138,368	1.050	84
10	DOE / NNSA Los Alamos Nat Lab	Roadrunner, IBM AMD + Cell GPU + IB	USA	122,400	1.04	76

November 2011: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx + custom	Japan	705,024	10.5	93	12.7	830
2	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel + Nvidia GPU + custom	China	186,368	2.57	55	4.04	636
3	DOE / OS Oak Ridge Nat Lab	Jaguar, Cray AMD + custom	USA	224,162	1.76	75	7.0	251
4	Nat. Supercomputer Center in Shenzhen	Nebulae, Dawning Intel + Nvidia GPU + IB	China	120,640	1.27	43	2.58	493
5	GSIC Center, Tokyo Institute of Technology	Tusbame 2.0, HP Intel + Nvidia GPU + IB	Japan	73,278	1.19	52	1.40	850
6	DOE / NNSA LANL & SNL	Cielo, Cray AMD + custom	USA	142,272	1.11	81	3.98	279
7	NASA Ames Research Center/NAS	Plelades SGI Altix ICE 8200EX/8400EX + IB	USA	111,104	1.09	83	4.10	265
8	DOE / OS Lawrence Berkeley Nat Lab	Hopper, Cray AMD + custom	USA	153,408	1.054	82	2.91	362
9	Commissariat a l'Energie Atomique (CEA)	Tera-10, Bull Intel + IB	France	138,368	1.050	84	4.59	229
10	DOE / NNSA Los Alamos Nat Lab	Roadrunner, IBM AMD + Cell GPU + IB	USA	122,400	1.04	76	2.35	446
500	IT Service	IBM Cluster, Intel + GigE	USA	7,236	.051	53		

Quiz: How Many of the Top500 systems use GPUs?

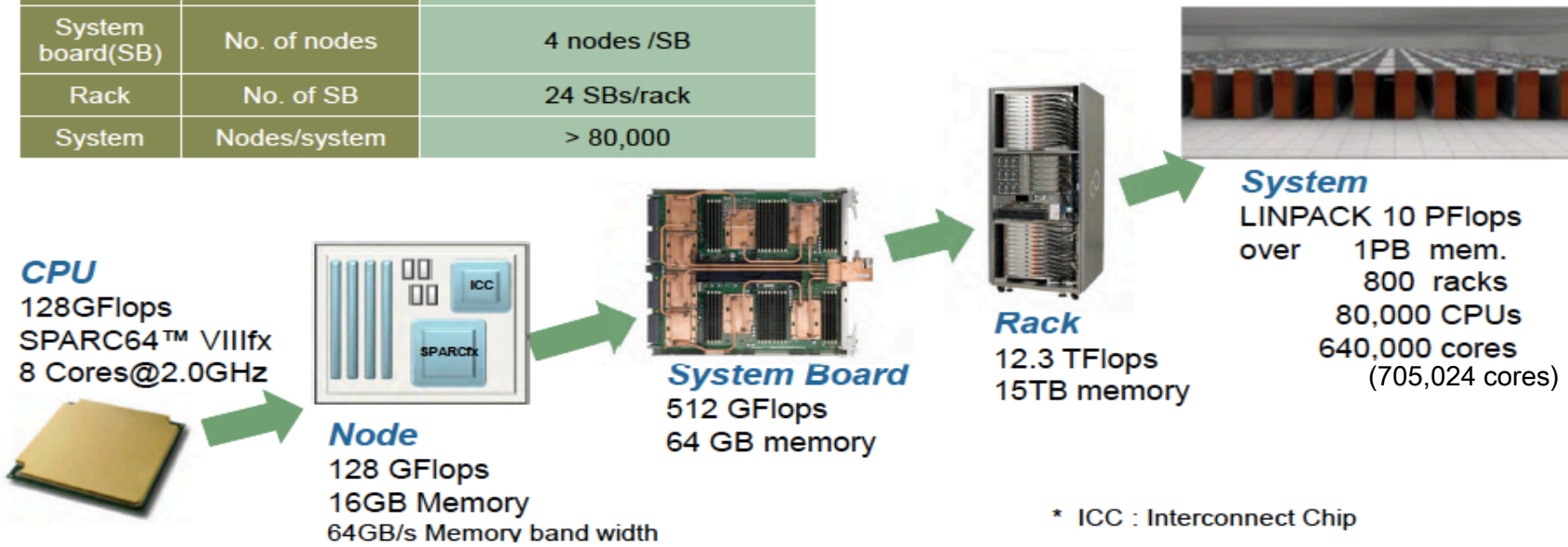
Japanese K Computer

K Computer > Sum(#2 : #8)
~ 2.5X #2

K computer Specifications

CPU (SPARC64 VIIIfx)	Cores/Node	8 cores (@2GHz)
	Performance	128GFlops
	Architecture	SPARC V9 + HPC extension
	Cache	L1(I/D) Cache : 32KB/32KB L2 Cache : 6MB
	Power	58W (typ. 30 C)
	Mem. bandwidth	64GB/s.
Node	Configuration	1 CPU / Node
	Memory capacity	16GB (2GB/core)
System board(SB)	No. of nodes	4 nodes /SB
Rack	No. of SB	24 SBs/rack
System	Nodes/system	> 80,000

Inter-connect	Topology	6D Mesh/Torus
	Performance	5GB/s. for each link
	No. of link	10 links/ node
	Additional feature	H/W barrier, reduction
	Architecture	Routing chip structure (no outside switch box)
Cooling	CPU, ICC*	Direct water cooling
	Other parts	Air cooling



07 Linpack run with 705,024 cores at 10.51 Pflop/s (88,128 CPUs), 12.7 MW; 29.5 hours
Fujitsu to have a 100 Pflop/s system in 2014

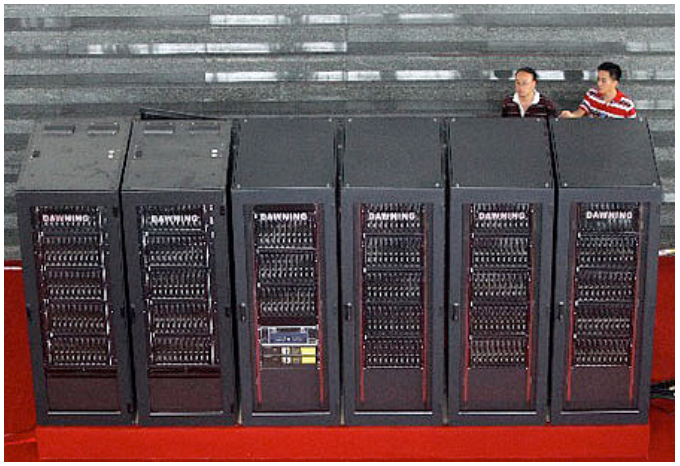
China's Very Aggressive Deployment of HPC

Absolute Counts	
US:	263
China:	75
Japan:	30
UK:	27
France:	23
Germany:	20

- China has 6 Pflops systems (4 based on GPUs)
  - 2-NUDT, Tianhe-1A, located in Tianjin

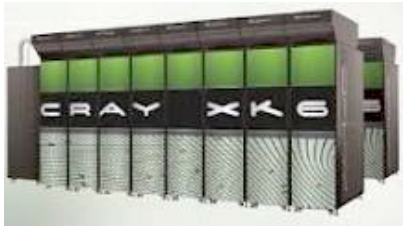
    Dual-Intel 6 core + Nvidia Fermi w/custom interconnect
    - Budget 600M RMB
      - MOST 200M RMB, Tianjin Government 400M RMB
  - CIT, Dawning 6000, Nebulea, located in Shenzhen

    Dual-Intel 6 core + Nvidia Fermi w/QDR Ifiniband
    - Budget 600M RMB
      - MOST 200M RMB, Shenzhen Government 400M RMB
  - Mole-8.5 Cluster/320x2 Intel QC Xeon E5520 2.26 Ghz + 320x6 Nvidia Tesla C2050/QDR Infiniband



10+ Pflop/s Systems Planned in the States

- DOE Funded, Titan at Oak Ridge Nat. Lab, Cray design w/AMD & Nvidia, XE6/XK6 hybrid
  - 20 Pflop/s, 2012
- DOE Funded, Sequoia at Lawrence Livermore Nat. Lab, IBM's BG/Q
  - 20 Pflop/s, 2012
- DOE Funded, BG/Q at Argonne National Lab, IBM's BG/Q
  - 10 Pflop/s, 2012
- NSF Funded, Blue Waters at U of Illinois UC, Cray design w/AMD & Nvidia, XE6/XK6 hybrid
  - 11.5 Pflop/s, 2012
- NSF Funded, U of Texas, Austin, Based on Dell/Intel MIC
  - 10 Pflop/s, 2013



Commodity plus Accelerator

Commodity

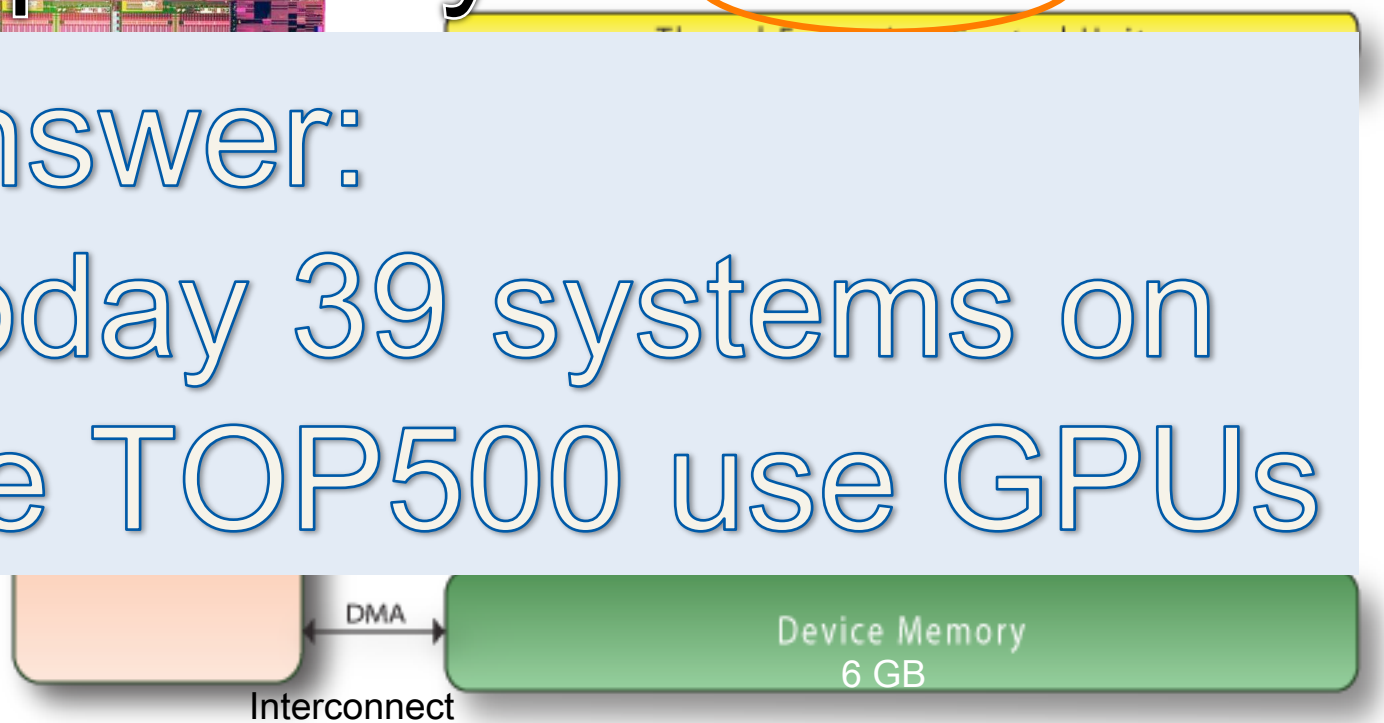
Intel Xeon
3 cores
3 GHz
8*4 ops/cycle
96 Gflop/s (DP)

Accelerator (GPU)

Nvidia C2050 "Fermi"
448 "Cuda cores"
1.15 GHz
448 ops/cycle
515 Gflop/s (DP)

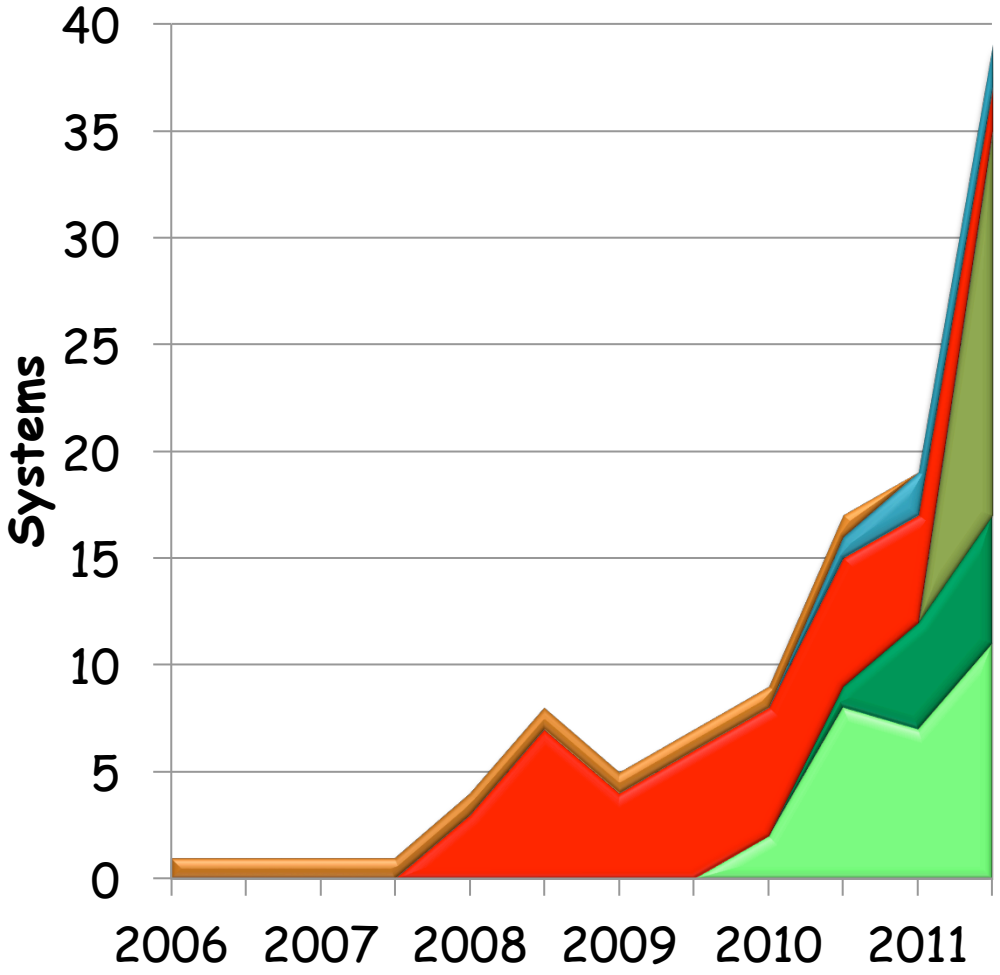
Quiz: How Many of the Top500 systems use GPUs?

Answer:
Today 39 systems on the TOP500 use GPUs



Interconnect
PCI-X 16 lane
64 Gb/s
1 GW/s

39 Accelerator Based Systems



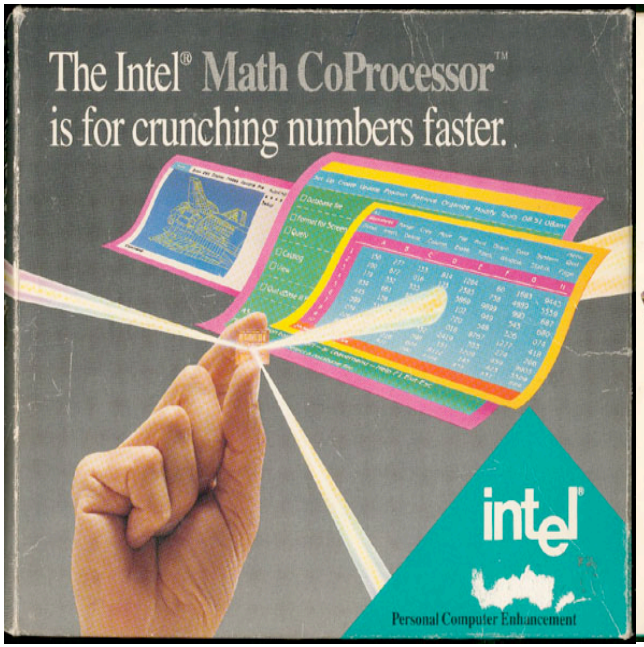
- Clearspeed CSX60022
- ATI GPU
- IBM PowerXCell 8i
- NVIDIA 2090
- NVIDIA 2070
- NVIDIA 2050

20 US	1 Italy
5 China	1 Poland
3 Japan	1 Spain
2 France	1 Switzerland
2 Germany	1 Russia
1 Australia	1 Taiwan

We Have Seen This Before

- Floating Point Systems FPS-164/ MAX Supercomputer (1976)
- Intel Math Co-processor (1980)
- Weitek Math Co-processor (1981)

1976



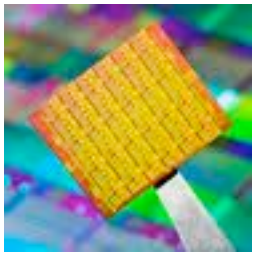
1980

Balance Between Data Movement and Floating point

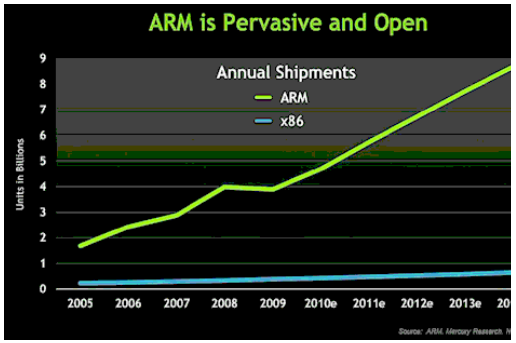
- FPS-164 and VAX (1976)
  - 11 Mflop/s; transfer rate 44 MB/s
  - Ratio of flops to bytes of data movement: 1 flop per 4 bytes transferred
- Nvidia Fermi and PCI-X to host
  - 500 Gflop/s; transfer rate 8 GB/s
  - Ratio of flops to bytes of data movement: 62 flops per 1 byte transferred
- Flop/s are cheap, so are provisioned in excess

Future Computer Systems

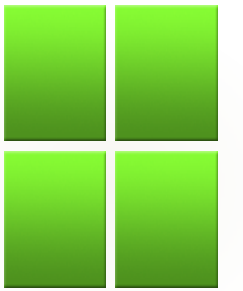
- Most likely be a hybrid design
  - Think standard multicore chips and accelerator (GPUs)
- Today accelerators are attached
- Next generation more integrated
- Intel's MIC architecture "Knights Ferry" and "Knights Corner" to come.
  - 48 x86 cores
- AMD's Fusion
  - Multicore with embedded graphics ATI
- Nvidia's Project Denver plans to develop an integrated chip using ARM architecture in 2013.

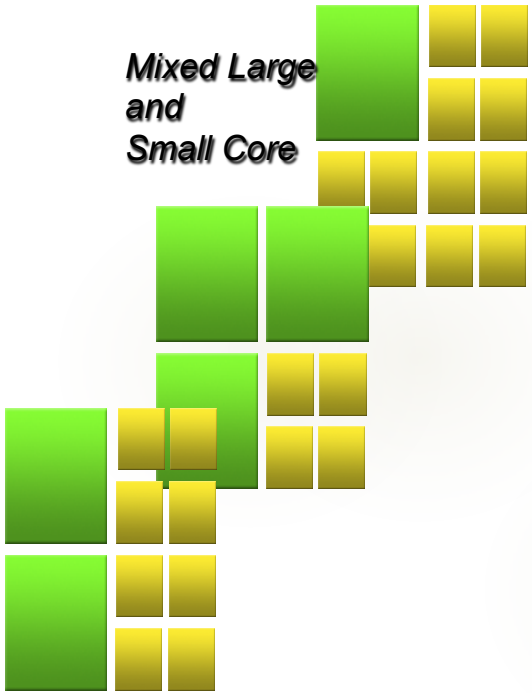


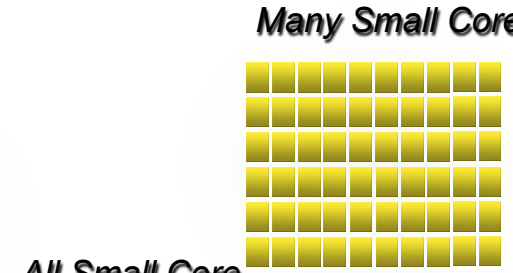
AMD
The future is fusion



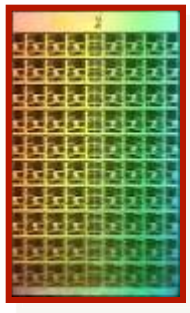
What's Next?

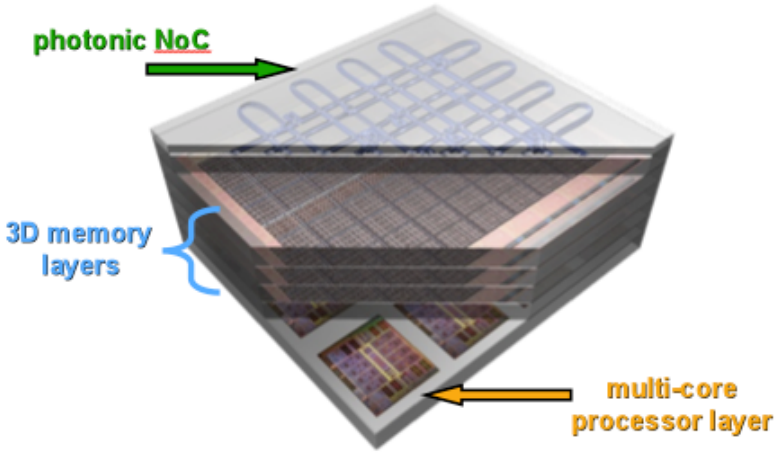






Many Floating-Point Cores





Different Classes of Chips

- Home
- Games / Graphics
- Business
- Scientific

The High Cost of Data Movement

•Flop/s or percentage of peak flop/s become much less relevant

Approximate power costs (in picoJoules)

	2011
DP FMADD flop	100 pJ
DP DRAM read	4800 pJ
Local Interconnect	7500 pJ
Cross System	9000 pJ

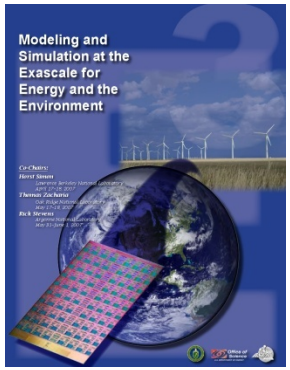
Source: John Shalf, LBNL

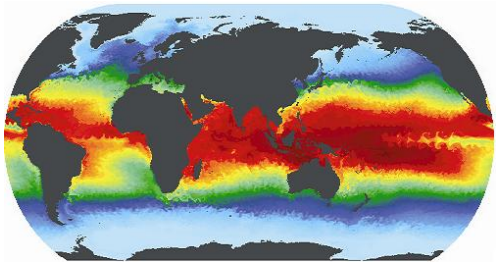
•Algorithms & Software: minimize data movement; perform more work per unit data movement.

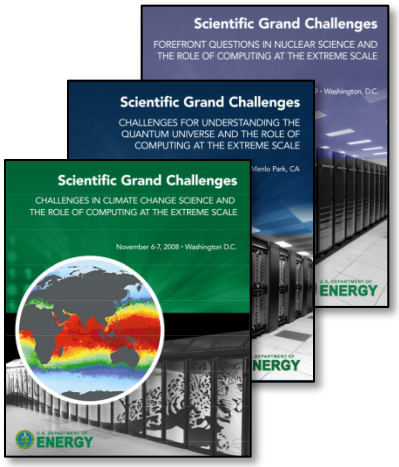
Broad Community Support and Development of the Exascale Initiative Since 2007

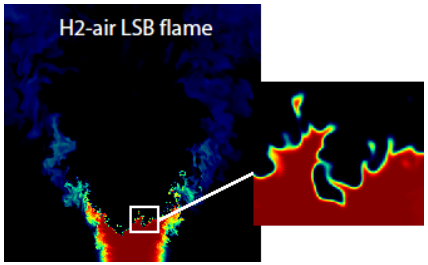
http://science.energy.gov/ascr/news-and-resources/program-documents/

- **Town Hall Meetings April-June 2007**
- **Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009**
  - **Climate Science (11/08)**
  - **High Energy Physics (12/08)**
  - **Nuclear Physics (1/09)**
  - **Fusion Energy (3/09)**
  - **Nuclear Energy (5/09)**
  - **Biology (8/09)**
  - **Material Science and Chemistry (8/09)**
  - **National Security (10/09)**
  - **Cross-cutting technologies (2/10)**
- **Exascale Steering Committee**
  - **"Denver" vendor NDA visits (8/09)**
  - **SC09 vendor feedback meetings**
  - **Extreme Architecture and Technology Workshop (12/09)**
- **International Exascale Software Project**
  - **Santa Fe, NM (4/09); Paris, France (6/09); Tsukuba, Japan (10/09); Oxford (4/10); Maui (10/10); San Francisco (4/11); Cologne (10/11)**

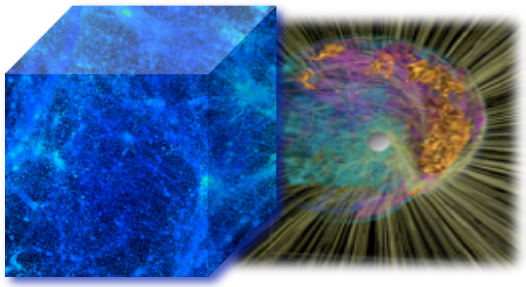






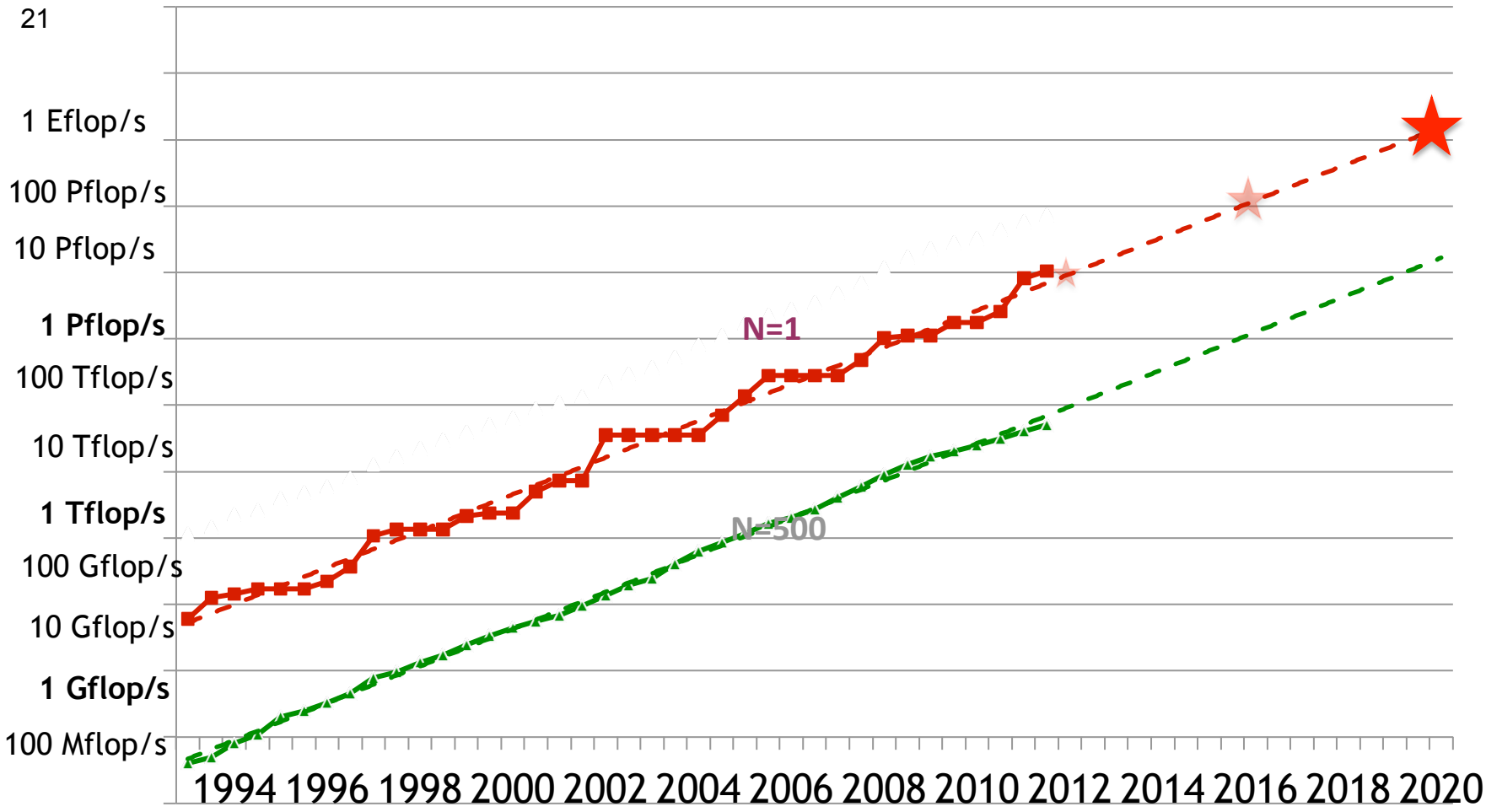


Mission Imperatives



Fundamental Science

Performance Development in Top500



Potential System Architecture

Systems	2011 K computer
System peak	**10.5 Pflop/s**
Power	**12.7 MW**
System memory	1.6 PB
Node performance	128 GF
Node memory BW	64 GB/s
Node concurrency	8
Total Node Interconnect BW	20 GB/s
System size (nodes)	88,124
Total concurrency	705,024
MTTI	days

Potential System Architecture with a cap of $200M and 20MW

Systems	2011 K computer	2019	Difference Today & 2019
System peak	**10.5 Pflop/s**	**1 Eflop/s**	**O(100)**
Power	**12.7 MW**	**~20 MW**	
System memory	1.6 PB	32 - 64 PB	O(10)
Node performance	128 GF	1,2 or 15TF	O(10) – O(100)
Node memory BW	64 GB/s	2 - 4TB/s	O(100)
Node concurrency	8	O(1k) or 10k	O(100) – O(1000)
Total Node Interconnect BW	20 GB/s	200-400GB/s	O(10)
System size (nodes)	88,124	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	705,024	O(billion)	O(1,000)
MTTI	days	O(1 day)	- O(10)

Major Changes to Software & Algorithms

- **Must rethink the design of our algorithms and software**
  - **Another disruptive technology**
    - Similar to what happened with cluster computing and message passing
  - **Rethink and rewrite the applications, algorithms, and software**
  - **Data movement is expense**
  - **Flop/s are cheap, so are provisioned in excess**

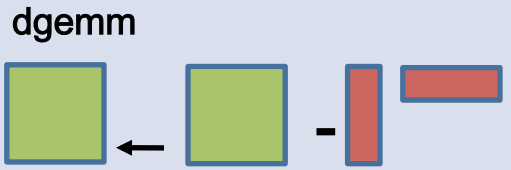
Critical Issues at Peta & Exascale for Algorithm and Software Design

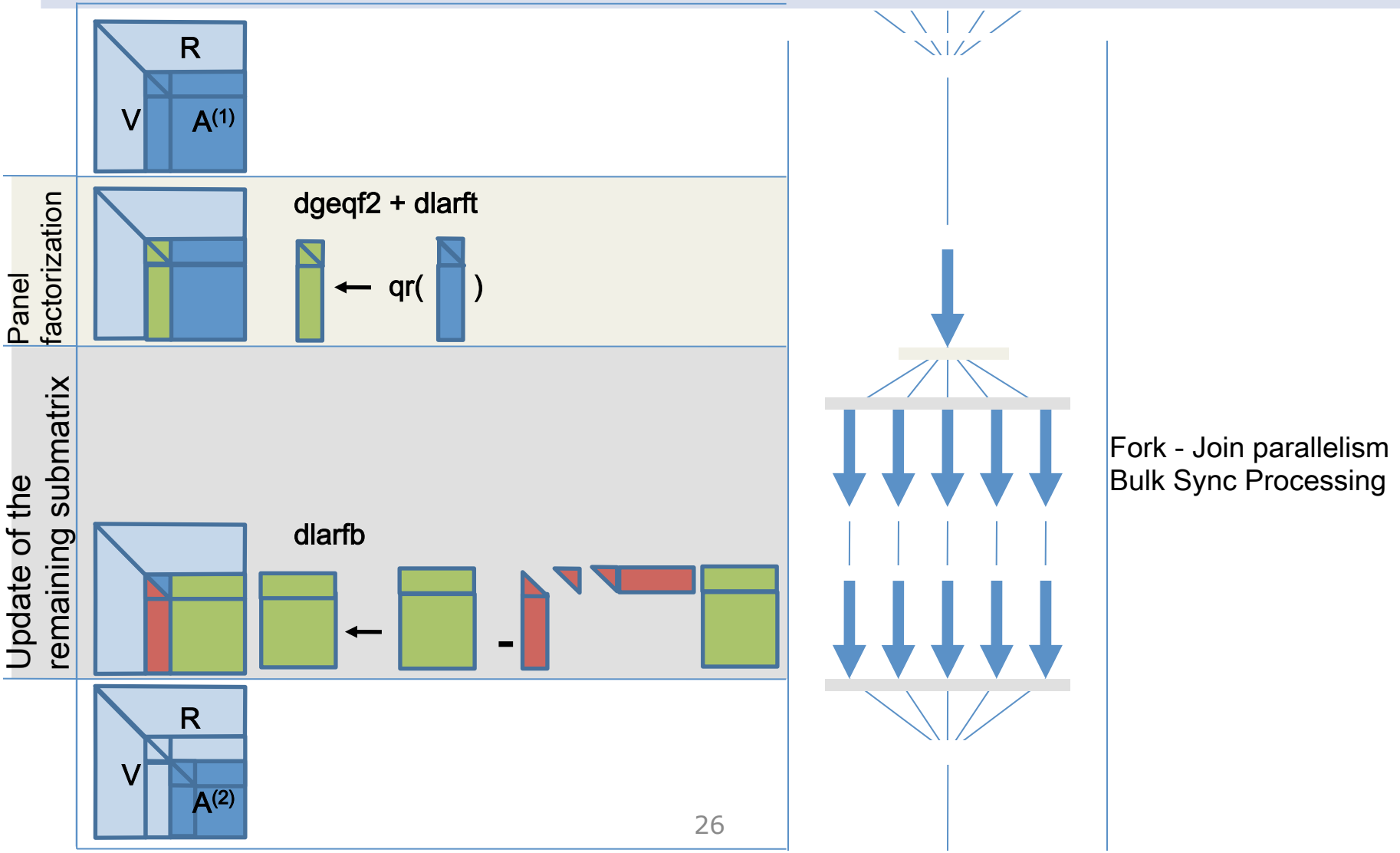
- Synchronization-reducing algorithms
  - Break Fork-Join model
- Communication-reducing algorithms
  - Use methods which have lower bound on communication
- Mixed precision methods
  - 2x speed of ops and 2x speed for data movement
- Autotuning
  - Today's machines are too complicated, build "smarts" into software to adapt to the hardware
- Fault resilient algorithms
  - Implement algorithms that can recover from failures/bit flips
- Reproducibility of results
  - Today we can't guarantee this. We understand the issues, but some of our "colleagues" have a hard time with this.

Parallelization of QR Factorization

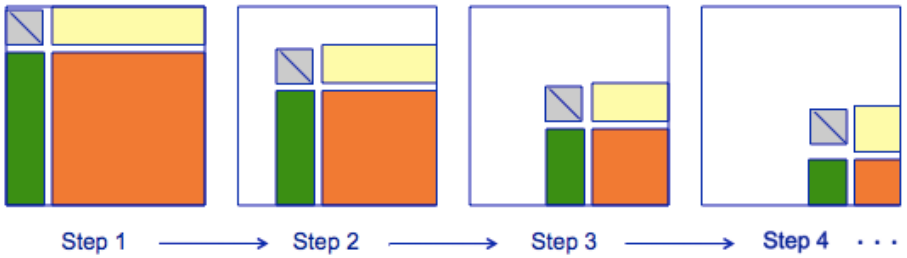
Parallelize the update:

- Easy and done in any reasonable software.
- This is the $2/3n^3$ term in the FLOPs count.
- Can be done "efficiently" with LAPACK+multithreaded BLAS

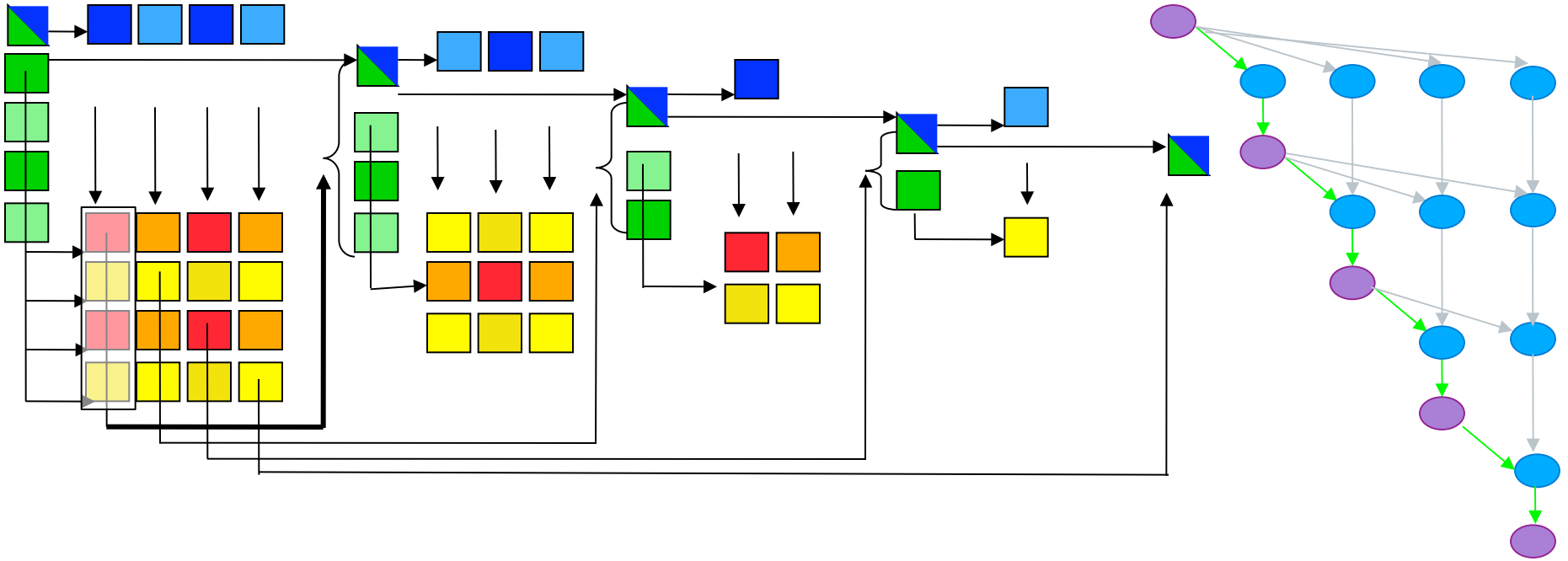




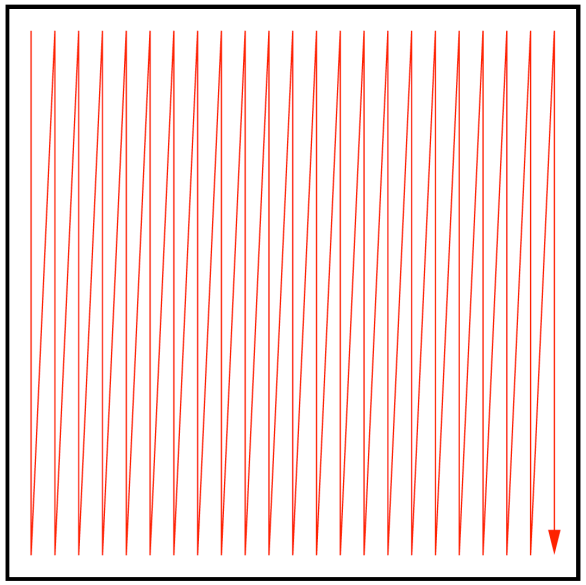
Parallel Tasks in LU/LL^T/QR

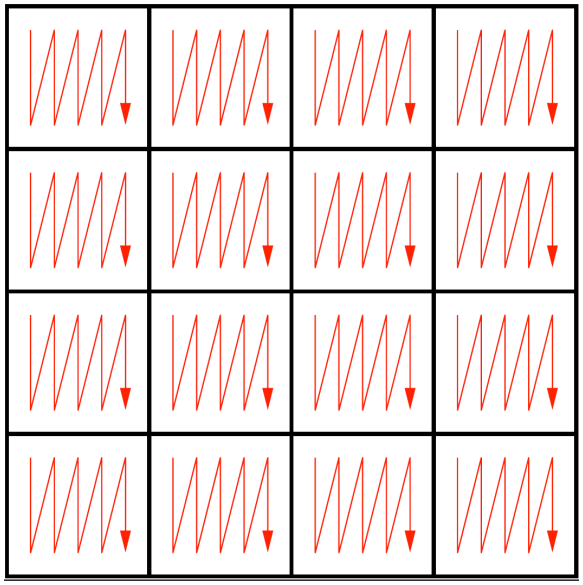


- Break into smaller tasks and remove dependencies



Data Layout is Critical

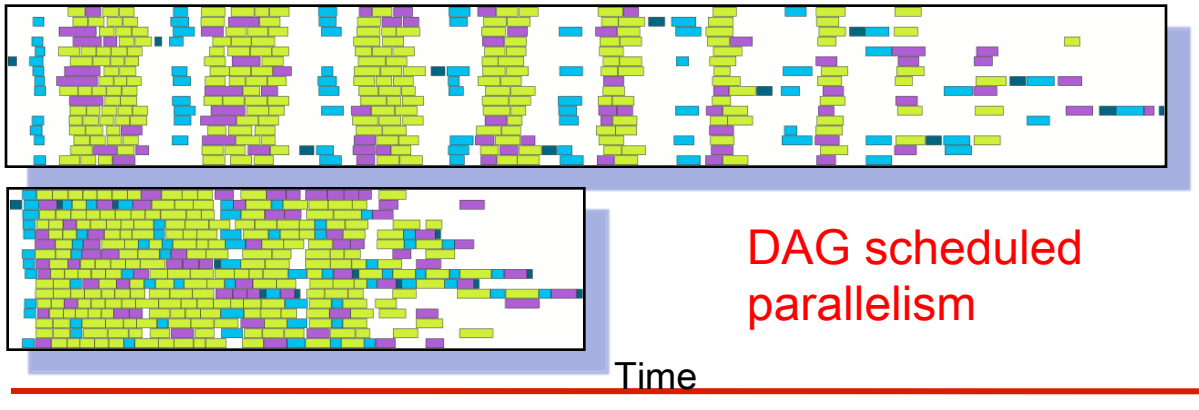


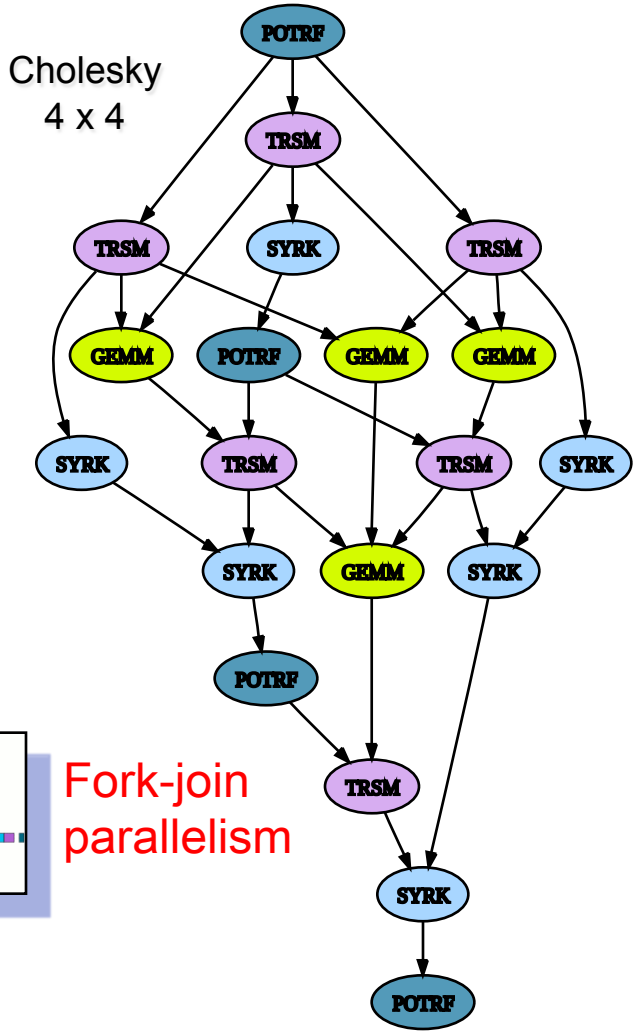


- Tile data layout where each data tile is contiguous in memory
- Decomposed into several fine-grained tasks, which better fit the memory of the small core caches

PLASMA: Parallel Linear Algebra s/w for Multicore Architectures

- **Objectives**
  - High utilization of each core
  - Scaling to large number of cores
  - Shared or distributed memory
- **Methodology**
  - Dynamic DAG scheduling (QUARK)
  - Explicit parallelism
  - Implicit communication
  - Fine granularity / block data layout
- **Arbitrary DAG with dynamic scheduling**

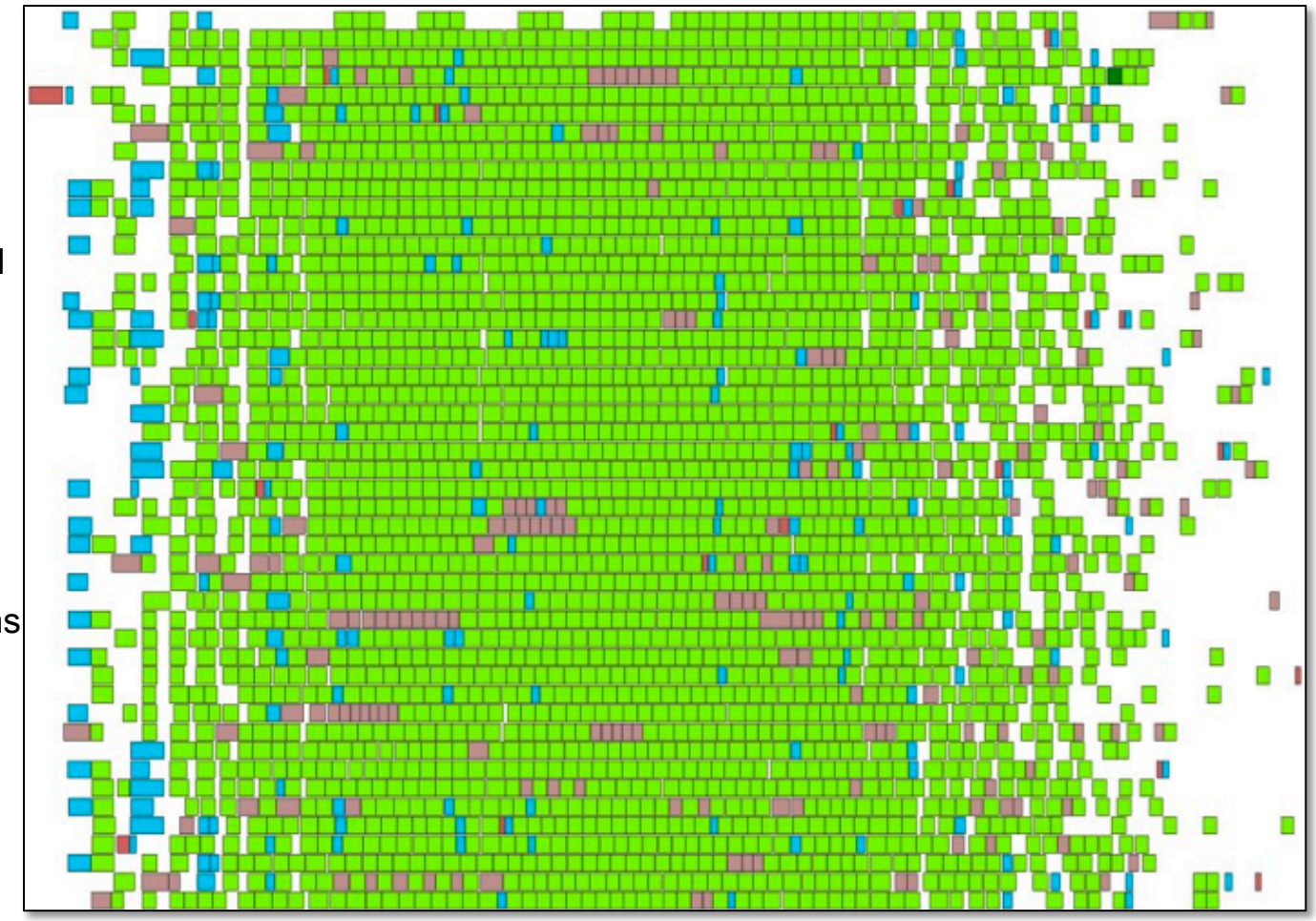




Synchronization Reducing Algorithms

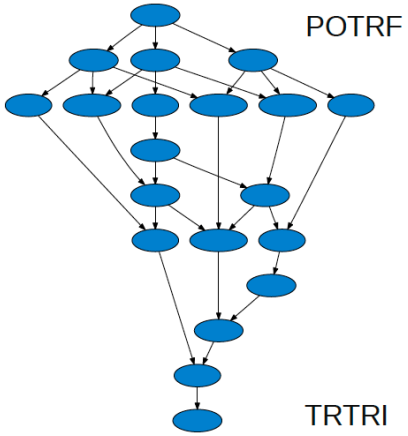
- Regular trace
- Factorization steps pipelined
- Stalling only due to natural load imbalance
- Dynamic
- Out of order execution
- Fine grain tasks
- Independent block operations

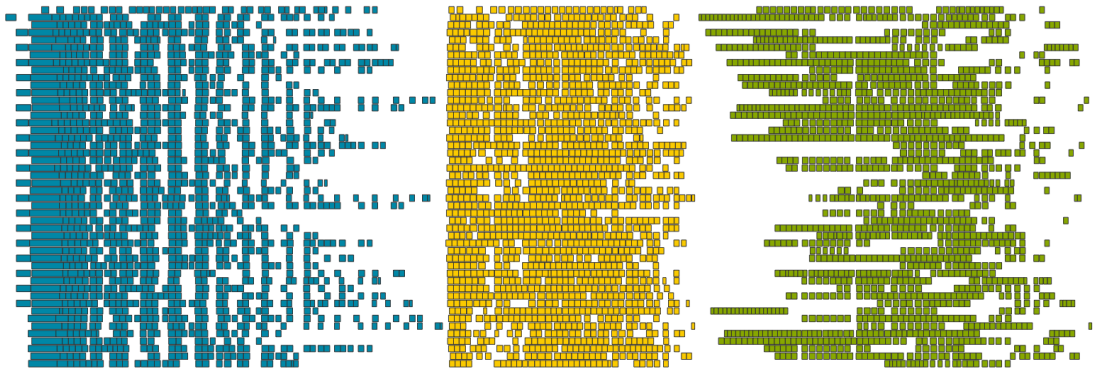
The colored area over the rectangle is the efficiency

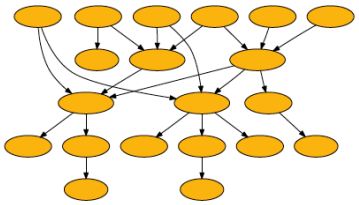


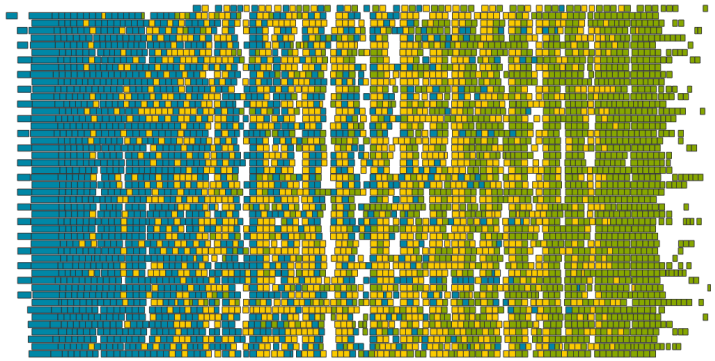
Tile QR factorization; Matrix size 4000x4000, Tile size 200
8-socket, 6-core (48 cores total) AMD Istanbul 2.8 GHz

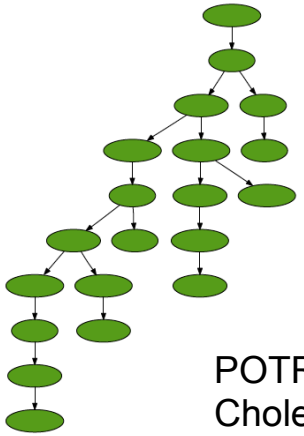
Pipelining: Cholesky Inversion
3 Steps: Factor, Invert L, Multiply L's





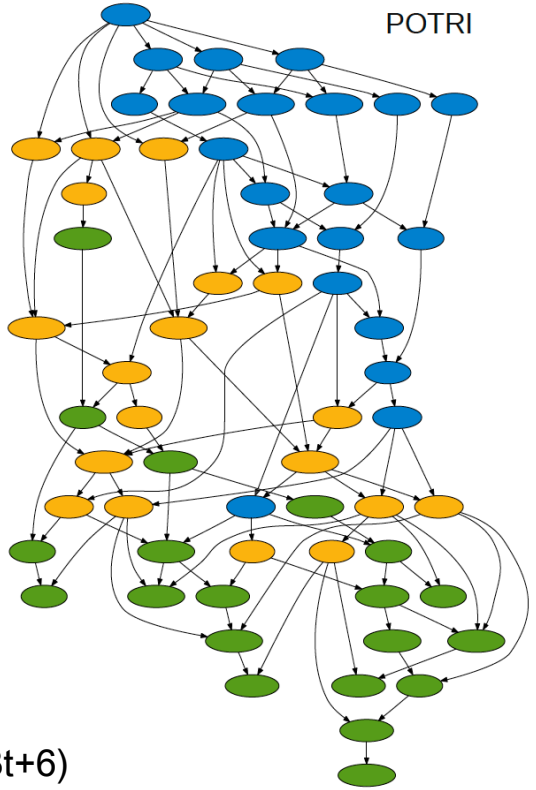






48 cores
POTRF, TRTRI and LAUUM.
The matrix is 4000 x 4000,tile size is 200 x 200,

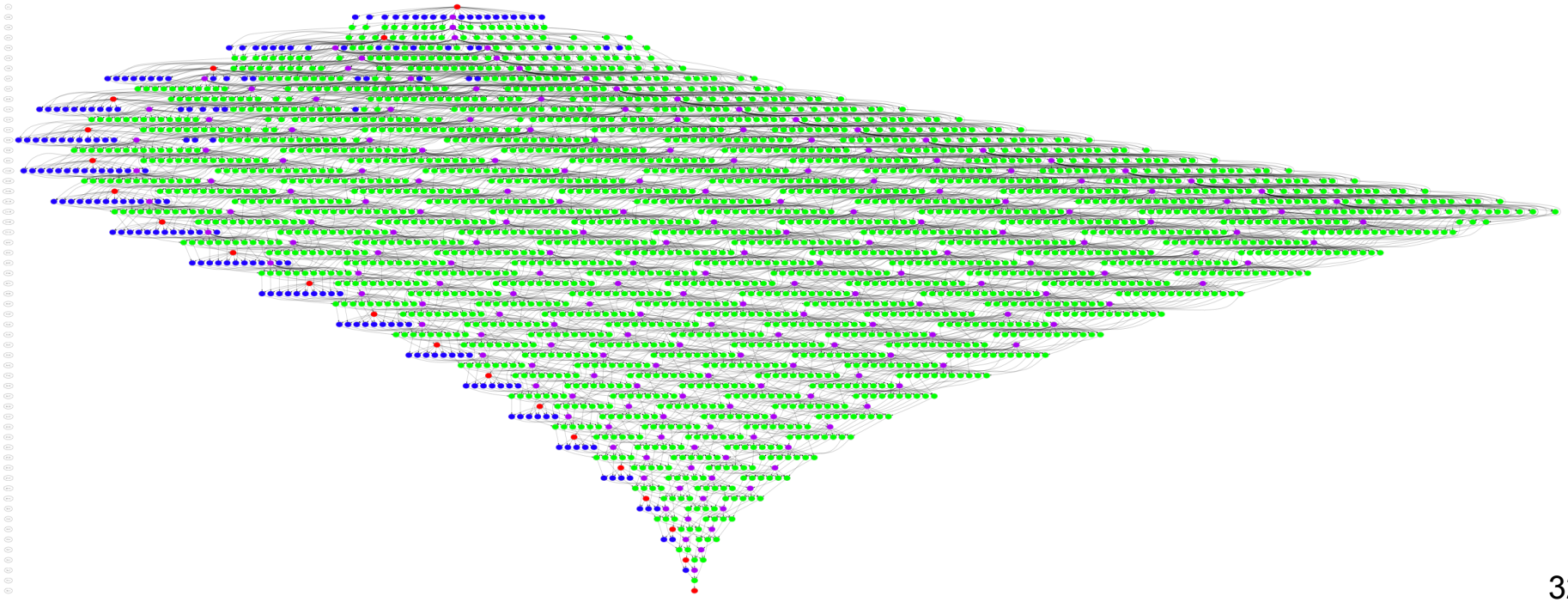
POTRF+TRTRI+LAUUM: 25 (7t-3)
Cholesky Factorization alone: 3t-2



Pipelined: 18 (3t+6)

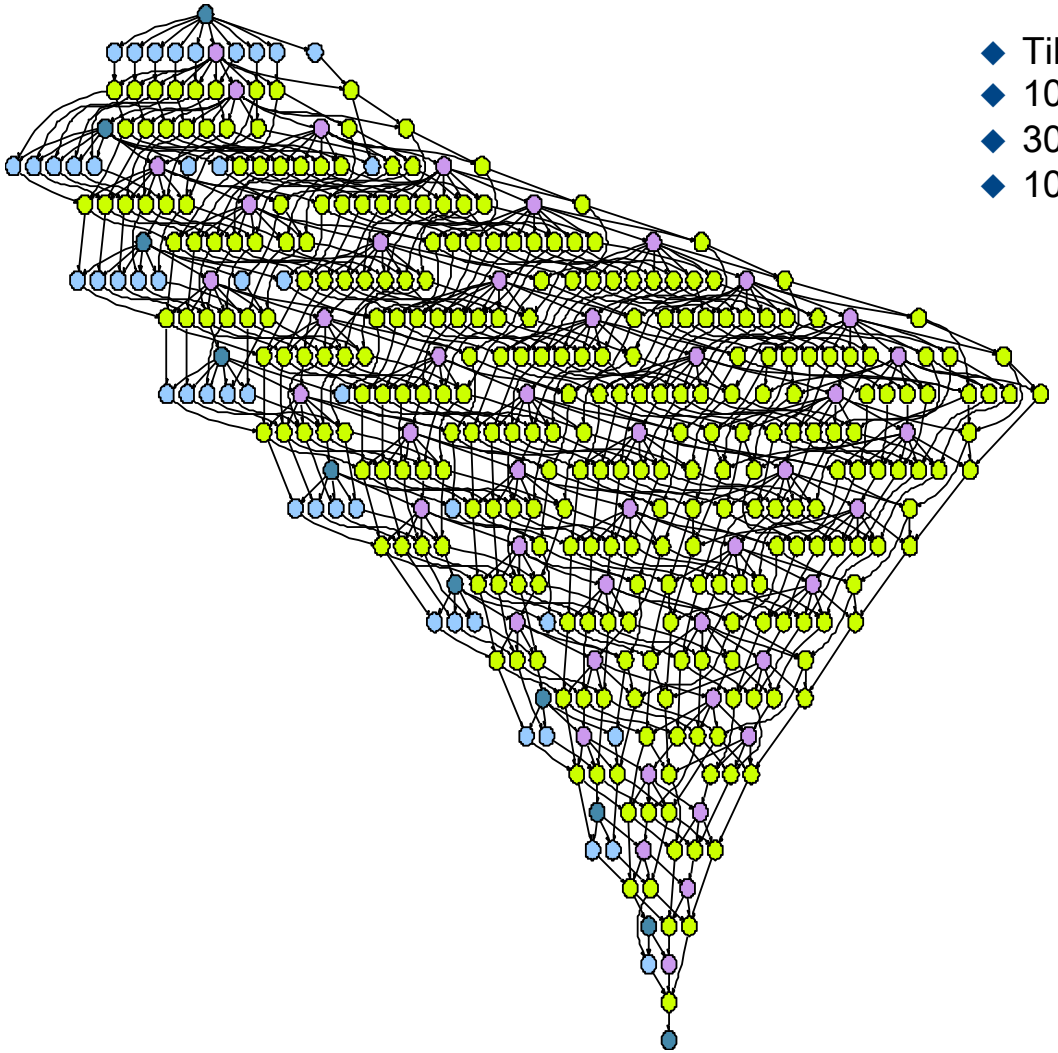
Big DAGs: No Global Critical Path

- DAGs get very big, very fast
  - So windows of active tasks are used; this means no global critical path
  - Matrix of NBxNB tiles; NB^3 operation
    - NB=100 gives 1 million tasks



PLASMA Local Scheduling

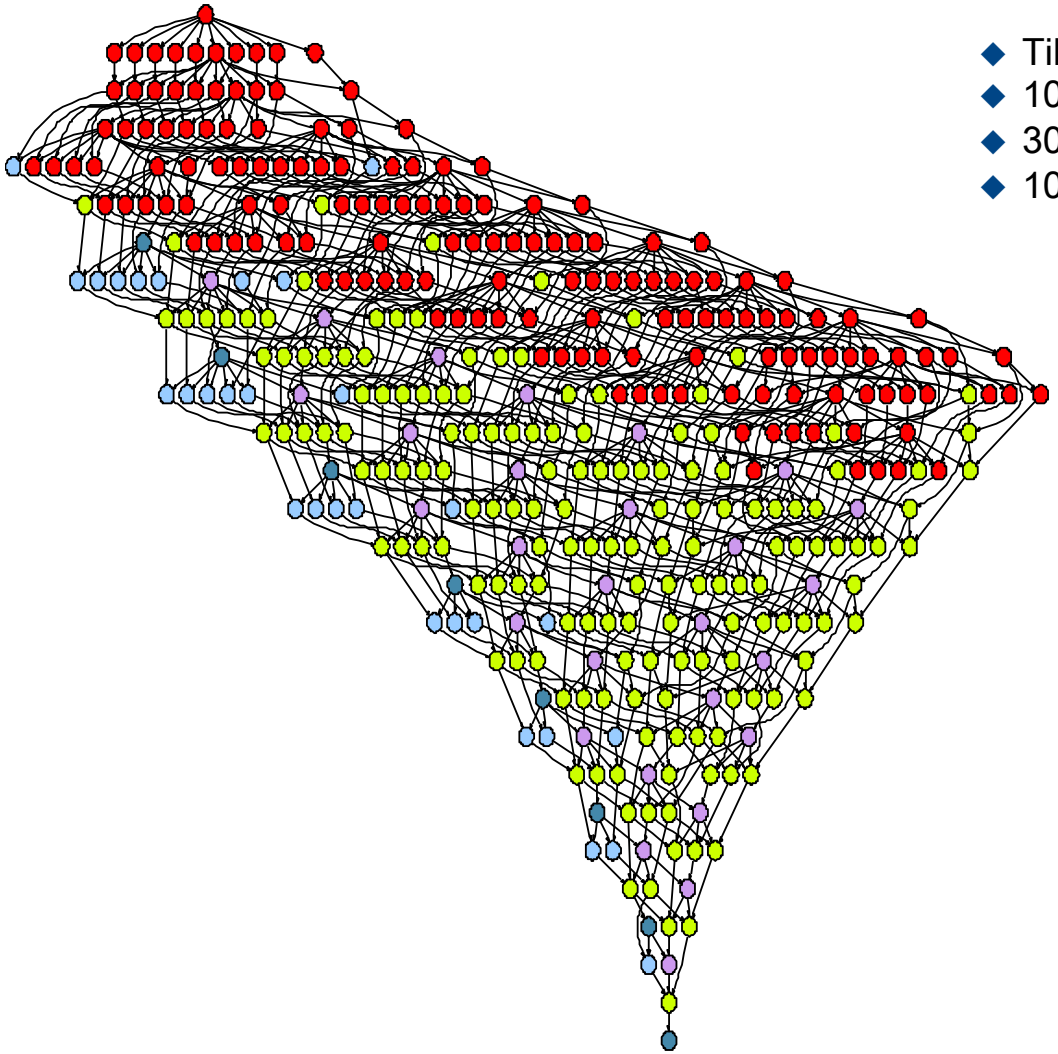
Dynamic Scheduling: Sliding Window



- Tile LU factorization
- 10 x 10 tiles
- 300 tasks
- 100 task window

PLASMA Local Scheduling

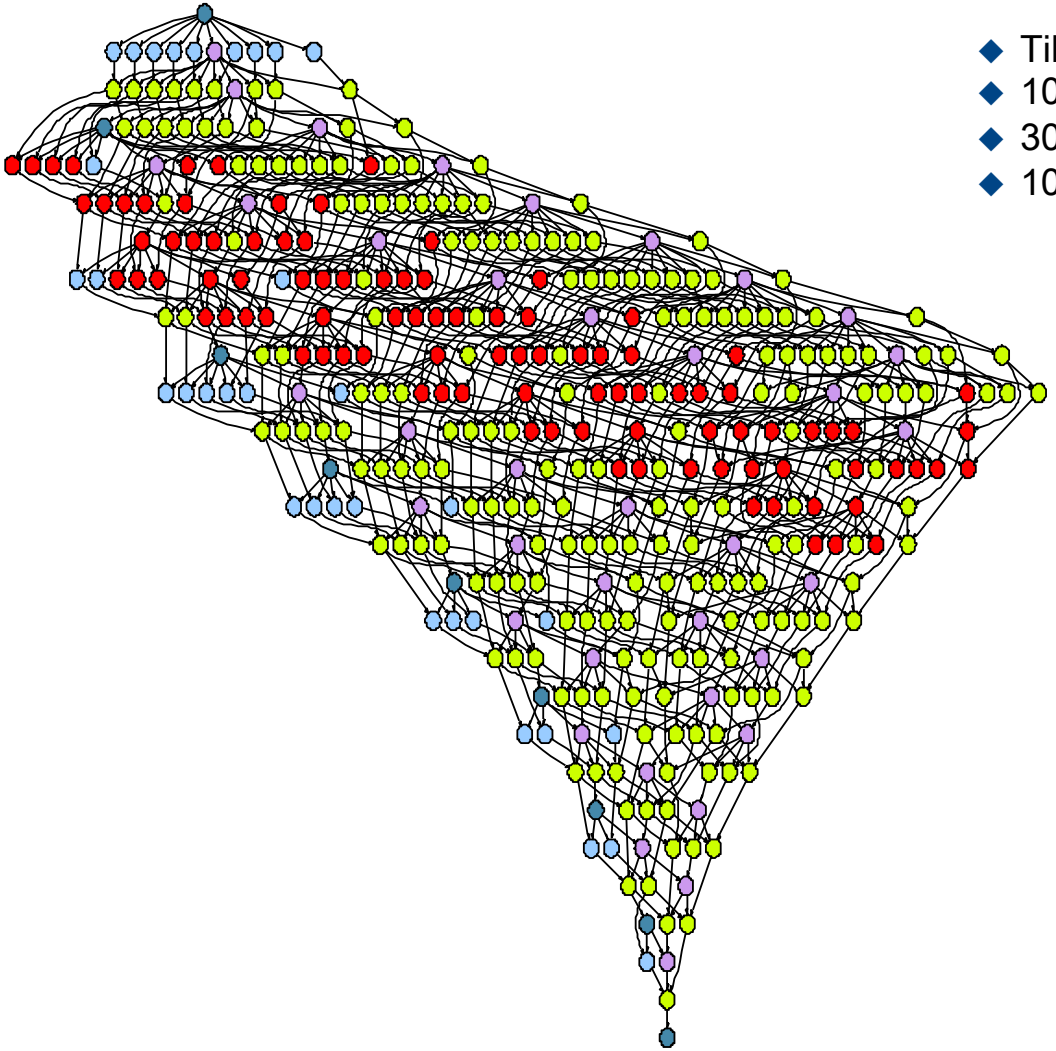
Dynamic Scheduling: Sliding Window



- Tile LU factorization
- 10 x 10 tiles
- 300 tasks
- 100 task window

PLASMA Local Scheduling

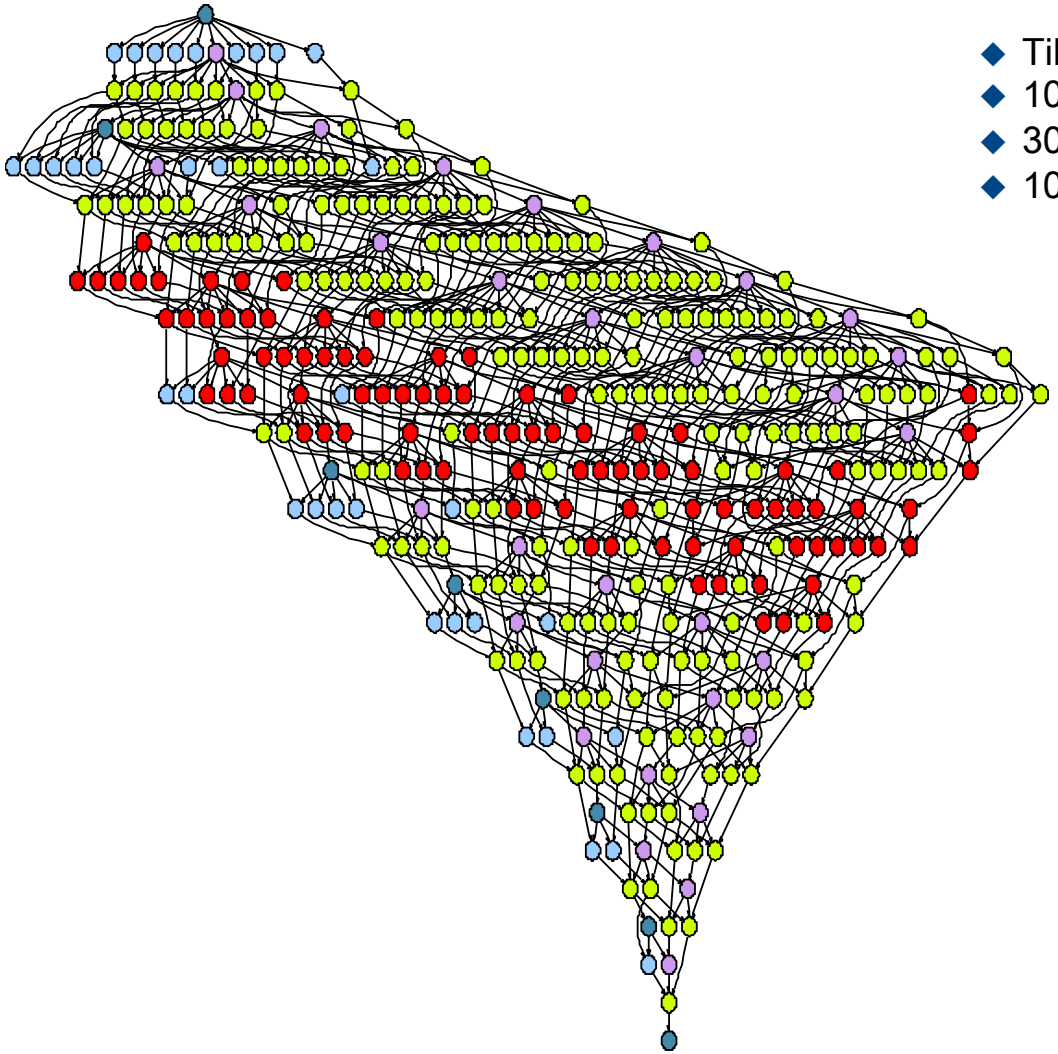
Dynamic Scheduling: Sliding Window



- Tile LU factorization
- 10 x 10 tiles
- 300 tasks
- 100 task window

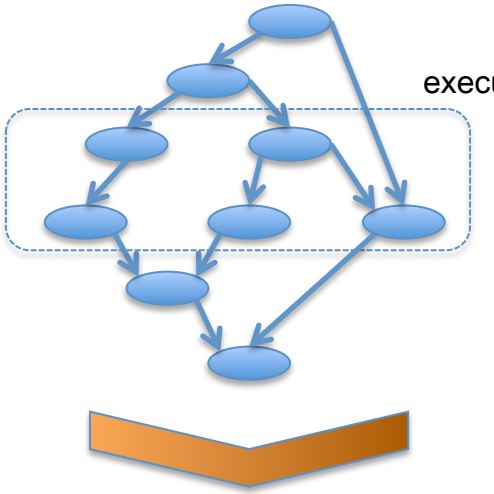
PLASMA Local Scheduling

Dynamic Scheduling: Sliding Window



- Tile LU factorization
- 10 x 10 tiles
- 300 tasks
- 100 task window

PLASMA
(On Node)



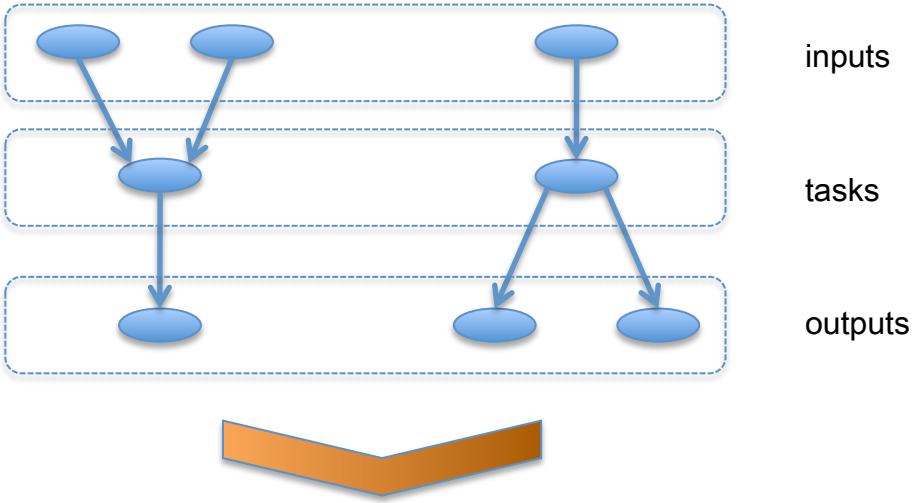
QUARK

Number of tasks in DAG:

$O(n^3)$

Cholesky: 1/3 n^3
LU: 2/3 n^3
QR: 4/3 n^3

DPLASMA
(Distributed System)



DAGuE

Number of tasks in parameterized DAG:

O(1)

Cholesky: 4 (POTRF, SYRK, GEMM, TRSM)
LU: 4 (GETRF, GESSM, TSTRF, SSSSM)
QR: 4 (GEQRT, LARFB, TSQRT, SSRFB)

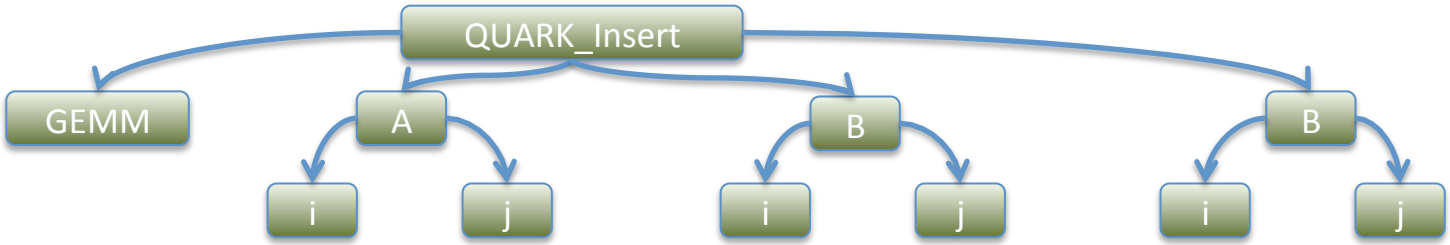
DAG: Conceptualized & Parameterized

small enough to store on each core in every node = Scalable

Start with PLASMA

```
for i,j = 0..N
   QUARK_Insert( GEMM,  A[i, j],INPUT,   B[j, i],INPUT,  C[i,i],INOUT )
   QUARK_Insert( TRSM,  A[i, j],INPUT,   B[j, i],INOUT )
```

Parse the C source code to Abstract Syntax Tree



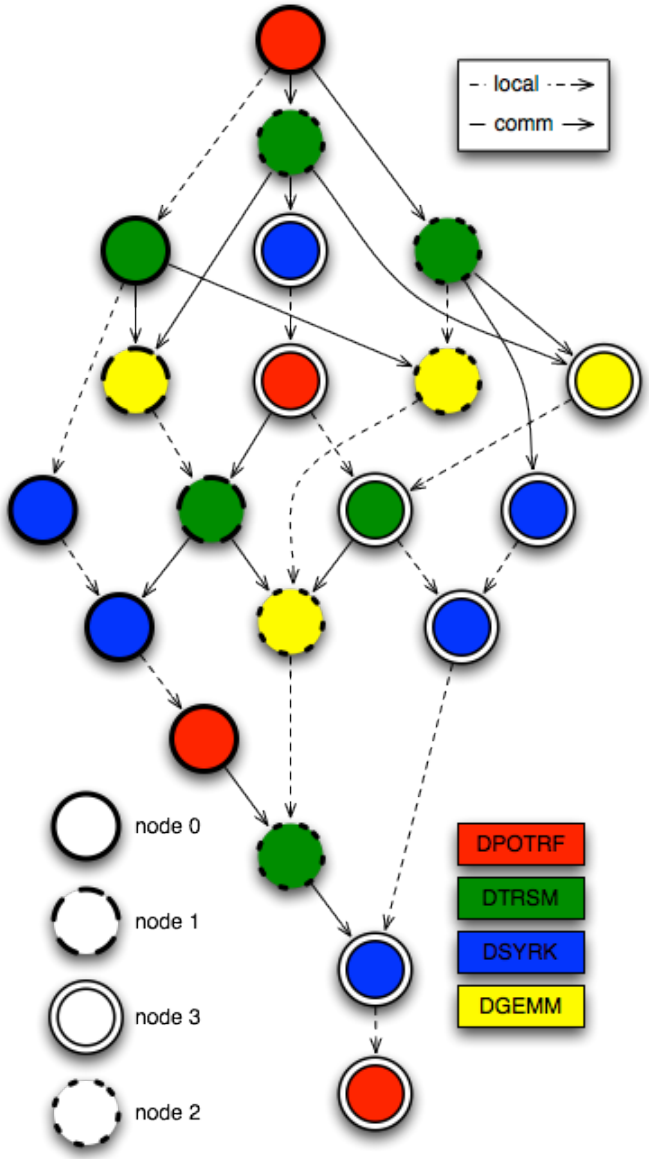
Analyze dependencies with Omega Test

```
{ 1 < i < N : GEMM(i, j) => TRSM(j) }
```

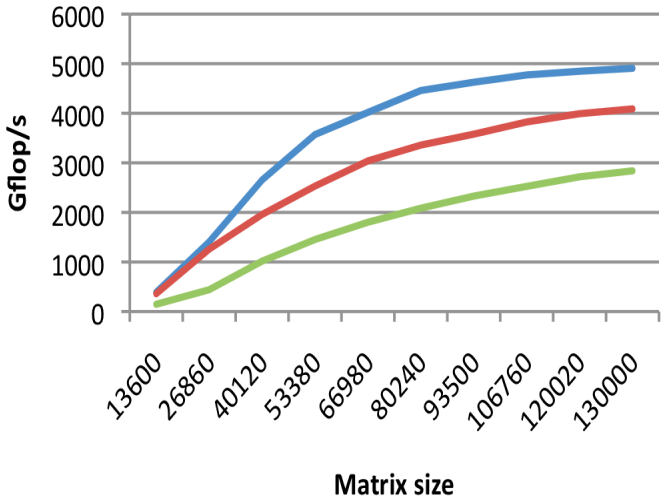
Loops & array references have to be affine

Generate Code which has the Parameterized DAG

Example: Cholesky 4x4



- RT is using the symbolic information from the compiler to make scheduling, message passing, & RT decisions
- Data distribution: regular, irregular
- Task priorities
- No left looking or right looking, more adaptive or opportunistic

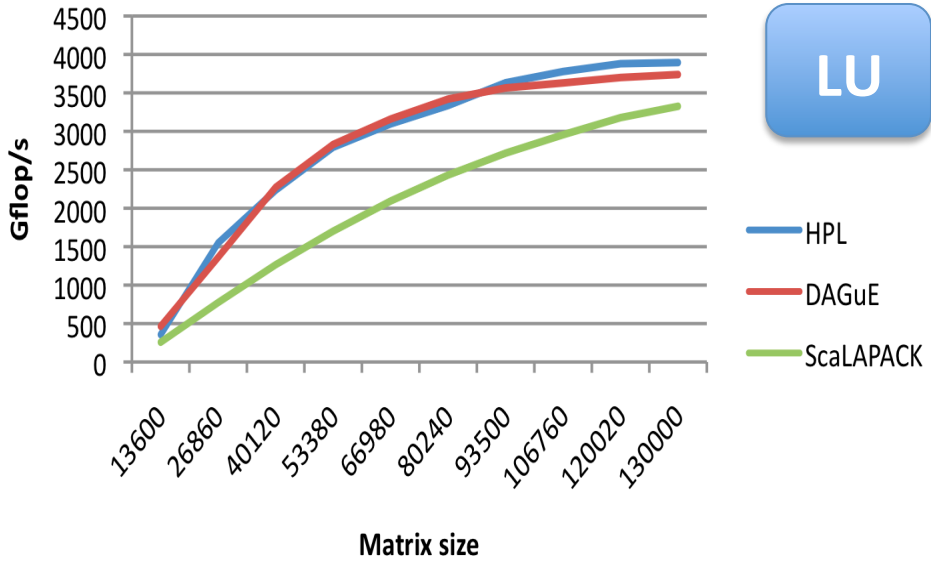


Cholesky

DAGuE
DSBP
ScaLAPACK

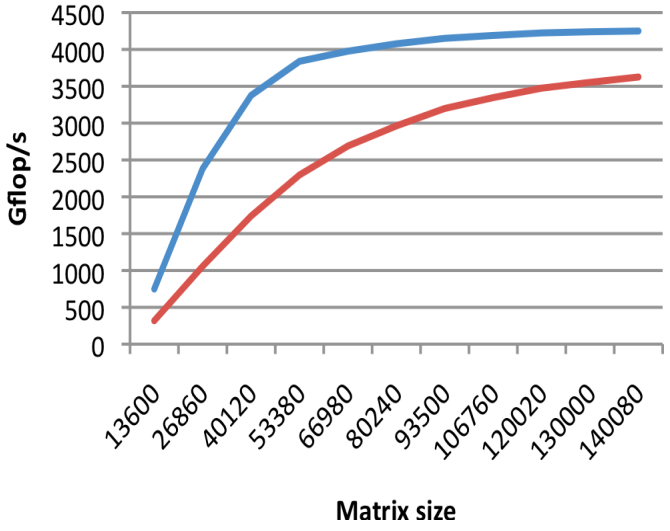
DSBP = Distributed Square Block Packed

81 nodes
Dual socket nodes
Quad core Xeon L5420
Total 648 cores at 2.5 GHz
ConnectX InfiniBand DDR 4x



LU

HPL
DAGuE
ScaLAPACK



QR

DAGuE
ScaLAPACK

Conclusions

- For the last decade or more, the research investment strategy has been overwhelmingly biased in favor of hardware.
- This strategy needs to be rebalanced - barriers to progress are increasingly on the software side.
- High Performance Ecosystem out of balance
  - Hardware, OS, Compilers, Software, Algorithms, Applications
    - No Moore's Law for software, algorithms and applications

Published in the January 2011 issue of The International Journal of High Performance Computing Applications

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

- *Alan Turing (1912 – 1954)*

- www.exascale.org