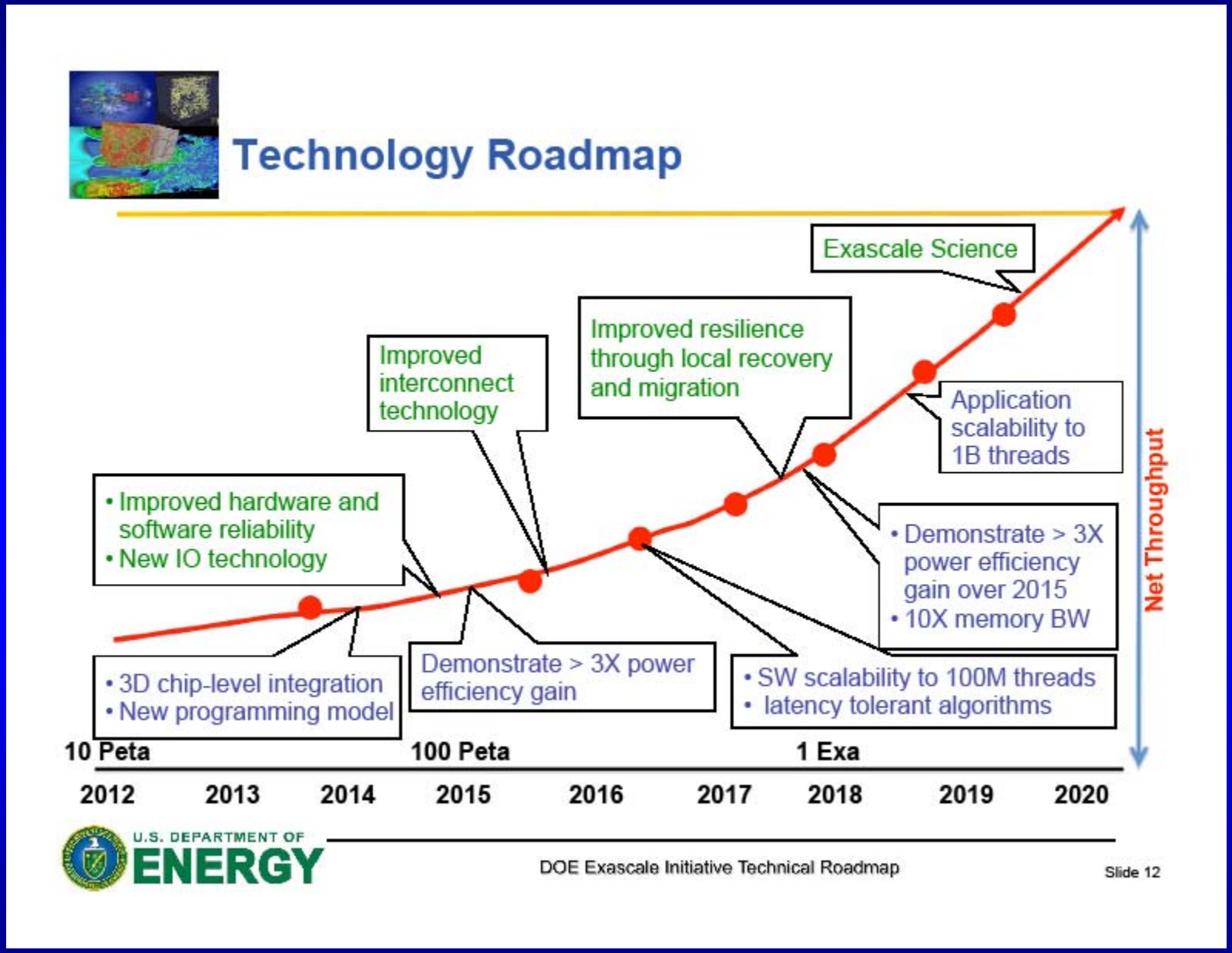
Co-Designing a COTS Re-configurable Exascale Computer

Steve Wallach
swallach”at”conveycomputer “dot”com
Convey <= Convex++

CONVEY computer

Acknowledgement

- Steve Poole and I had many emails and discussion on technology trends.
- We are both guilty of being overzealous architects


Technology Roadmap
Exascale Science
Improved resilience through local recovery and migration
Improved interconnect technology
Application scalability to 1B threads
• Improved hardware and software reliability
• New IO technology
• Demonstrate > 3X power efficiency gain over 2015
• 10X memory BW
• 3D chip-level integration
• New programming model
Demonstrate > 3X power efficiency gain
• SW scalability to 100M threads
• latency tolerant algorithms
Net Throughput
10 Peta
100 Peta
1 Exa
2012
2013
2014
2015
2016
2017
2018
2019
2020
U.S. DEPARTMENT OF ENERGY
DOE Exascale Initiative Technical Roadmap
Slide 12

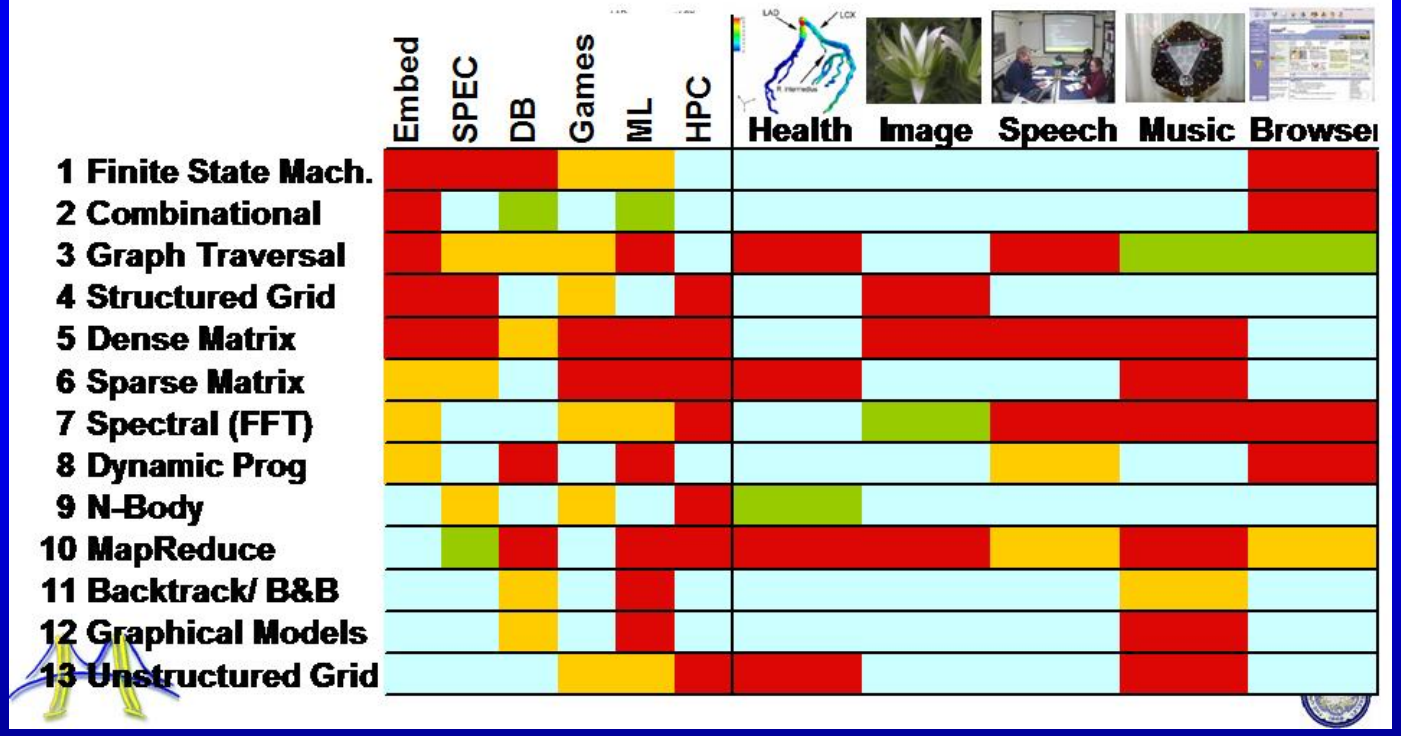
Systems	2009	2018 Swimlane 1	2018 SwimLane 2
System peak	**2 Peta**	**1 Exa**	
Power	**6 MW**	**~20 MW**	
System memory	0.3 PB	50 PB	
Node performance	125 GF	1TF	10TF
Interconnect Latency (for longest path)	1-5usec (limited by overhead at endpoints)	0.5usec (speed of light)	
Memory Latency	150-250 clock cycles (~70-100ns)	100 clock cycles (~50ns)	
Node memory BW	25 GB/s	0.4TB/s	4-5TB/s
Node concurrency	12	O(1k)	O(10k)
Total Node Interconnect BW	3.5 GB/s	200GB/s	2TB/s
System size (nodes)	18,700	O(1M)	O(100,000)
Total concurrency	225,000	O(100M)*10 for latency hiding	O(100M)*100 for latency hiding
Storage	15 PB	1000 PB (>10x system memory is min)	
IO	0.2 TB	60 TB/s	
MTABF (mean time between application failures)	Days	24 Hours	

Berkeley's 13 Motifs

"Motif/Dwarf" Popularity

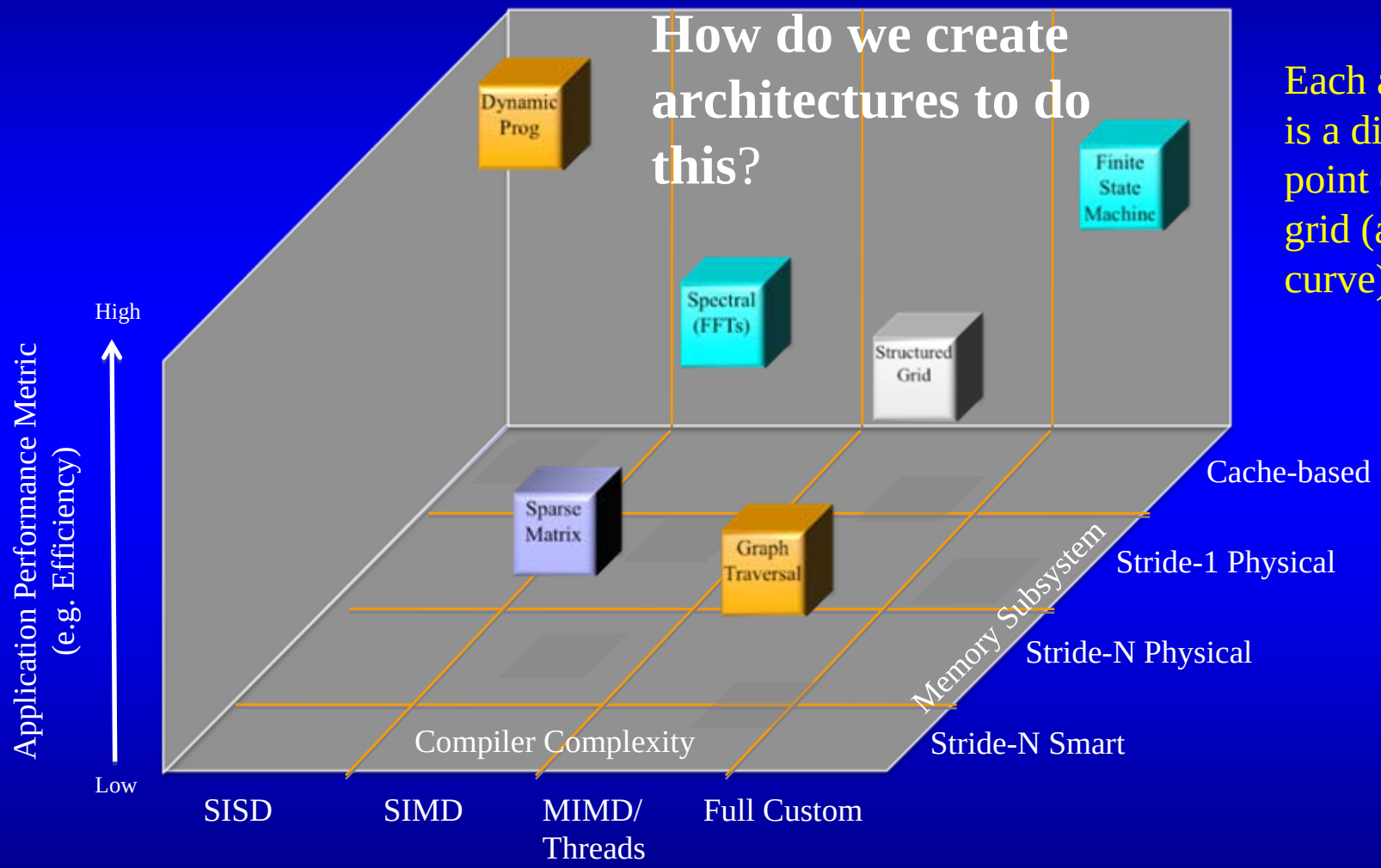
(Red Hot → Blue Cool)

- How do compelling apps relate to 13 motif/dwarfs?



http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-23.html

What does this all mean?



Each application is a different point on this 3D grid (actually a curve)

Take an Operation Research Method of Prediction

- Moore's Law
- Application Specific
  - Matrices
  - Matrix Arithmetic
- Hard IP for floating point
- Number of Arithmetic Engines
- System Arch and Programming model
  - Language directed design
- Resulting Analysis
  - Benefit
  - Mean and +/- sigma (if normal)

Bayesian Decision Theory

In a formal model the conclusions are derived from definitions and assumptions. . . .But with informal, verbal reasoning. . . one can argue until one is blue in the face . . . because
there is no criterion for deciding the soundness of an informal argument.

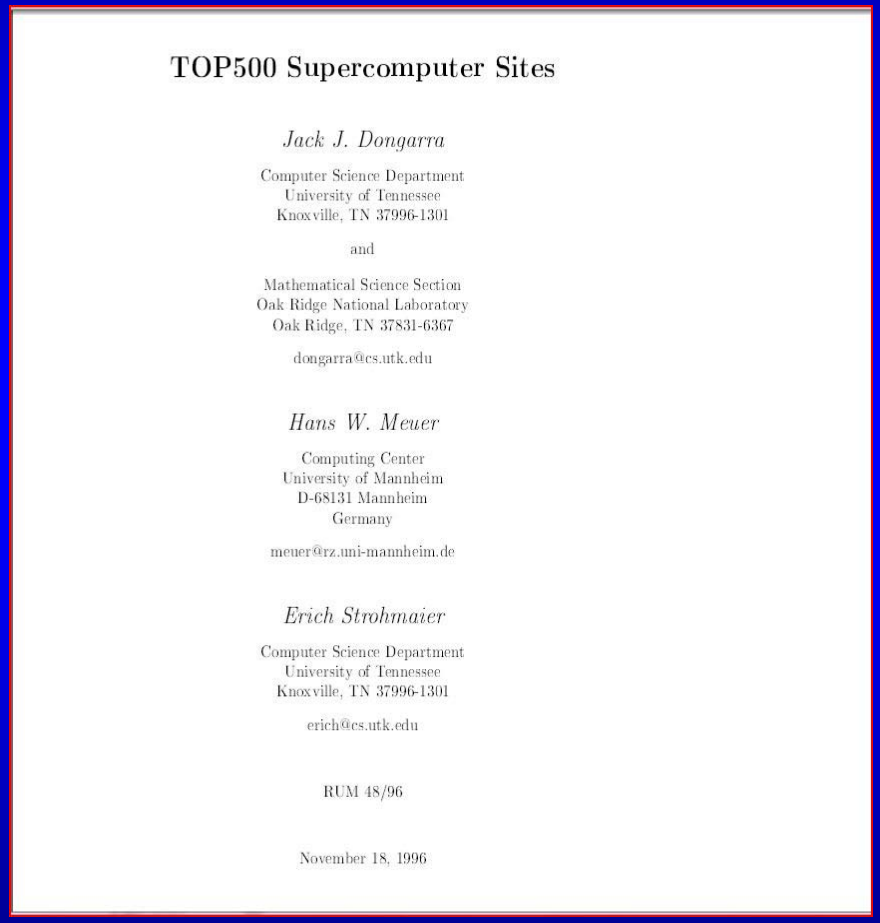
Robert
Aumann

Moore's Law

- Feature Set
  - Every 2 years twice the logic
  - Thus by 2020
    - 8 times the logic, same clock rate
  - Mean Factor of 7, sigma +/- 2
- Benefit
  - 8 times the performance, same clock rate, same internal architecture

Application Specific

- Feature Set
  - Matrix Engine
  - Mean Factor of 4 +4/- 1
    - Poisson
- June 2010 Linpack
  - Jaguar 75%
- June 1996 Linpack (% of peak)
  - NEC SX/4/20 = 95%
  - Fujitsu VPP/30 = 81%
- Benefit
  - 90% of Peak
  - Matrix Arithmetic's
  - Outer Loop Parallel/Inner Loop vector within ONE functional unit



TOP500 Supercomputer Sites

Jack J. Dongarra

Computer Science Department
University of Tennessee
Knoxville, TN 37996-1301

and

Mathematical Science Section
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6367

dongarra@cs.utk.edu

Hans W. Meuer

Computing Center
University of Mannheim
D-68131 Mannheim
Germany

meuer@rz.uni-mannheim.de

Erich Strohmaier

Computer Science Department
University of Tennessee
Knoxville, TN 37996-1301

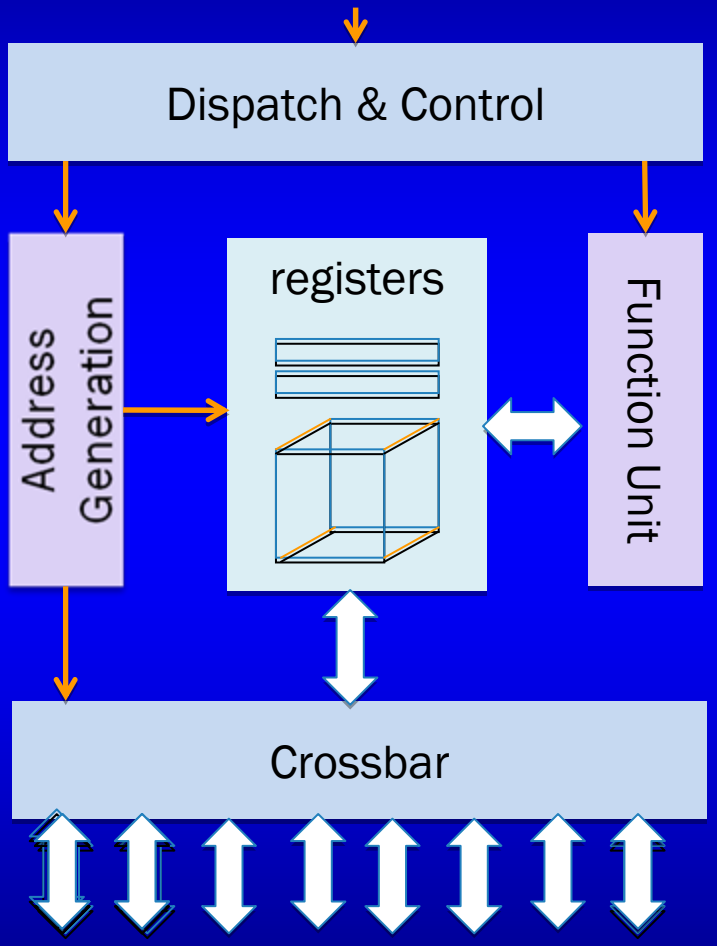
erich@cs.utk.edu

RUM 48/96

November 18, 1996

3D

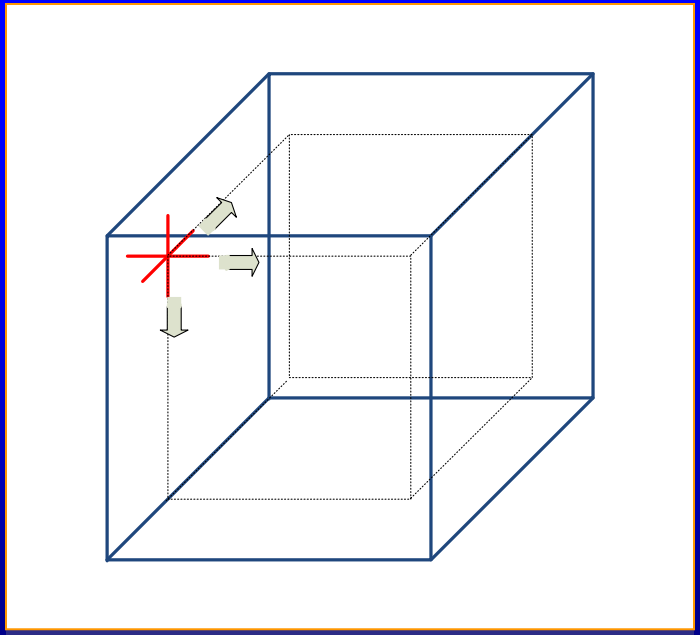
- Modular content
  - Datatypes: IEEE SP, IEEE DP, 32-bit integer, 64-bit integer, bit
  - Operations: simple, compound, reductions
  - Dimensionality: 1d, 2d, 3d
  - memory access: gather/scatter, strided, under mask, multidimensional
  - register access pattern: halo, sparse periodic access, under mask, multidimensional
- Number of function units determined by functionality selected



3D Finite Difference (3DFD) Personality

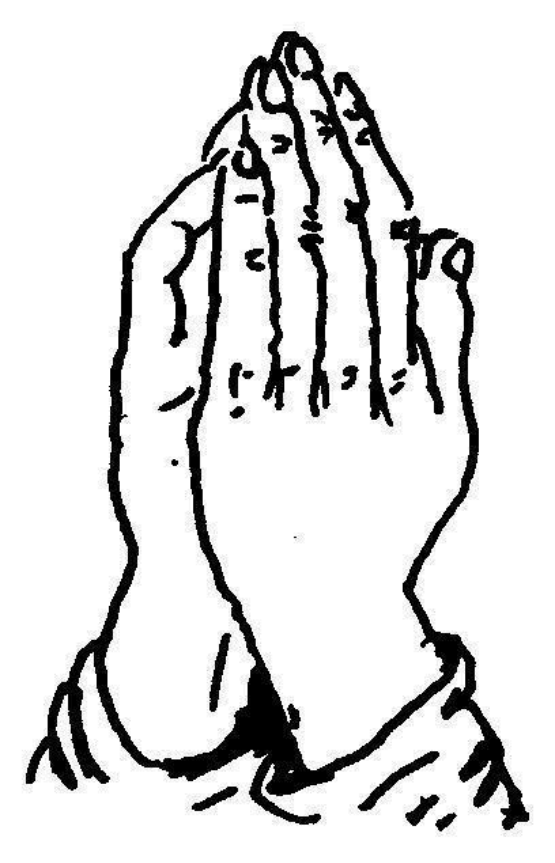
- designed for nearest neighbor operations on structured grids
  - maximizes data reuse
- reconfigurable "registers"
  - 1D (vector), 2D, and 3D modes
  - 8192 elements in a register
- operations on entire cubes
  - "add points to their neighbor to the left times a scalar" is a single instruction
  - up to 7 points away in any direction
- finite difference method for post-stack reverse-time migration

```
X(I,J,K) = S0*Y(I  ,J  ,K  )
         + S1*Y(I-1,J  ,K  )
         + S2*Y(I+1,J  ,K  )
         + S3*Y(I  ,J-1,K  )
         + S4*Y(I  ,J+1,K  )
         + S5*Y(I  ,J  ,K-1)
         + S6*Y(I  ,J  ,K+1)
```



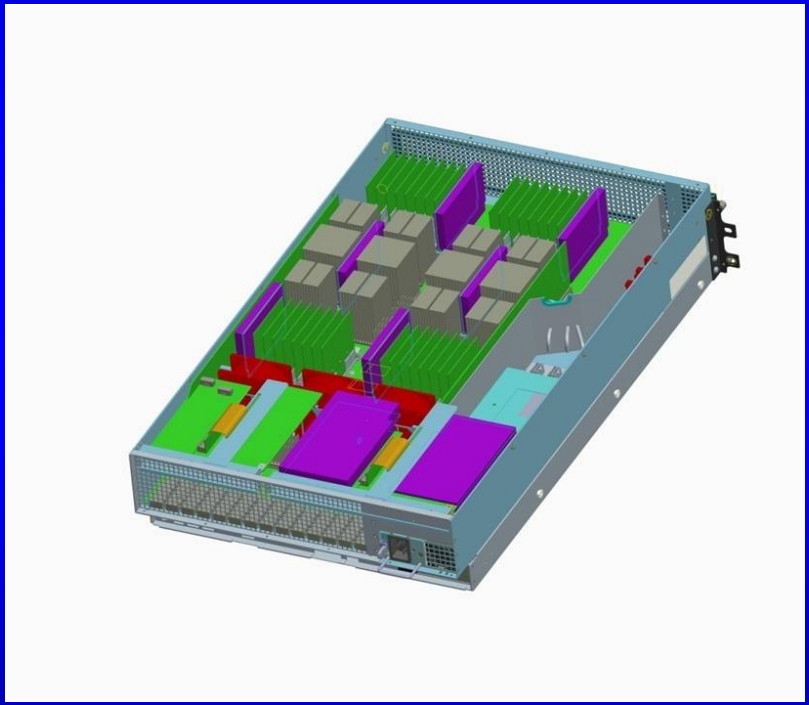
FPGA -Hard IP Floating Point

- Feature Set
  - By 2020
  - Fused Multiply Add – Mean Factor of 8 +/- 2
  - Reconfigurable IP Multiply and Add – Mean Factor 4 +/-1
  - Bigger fixed point DSP's – Mean Factor of 3
- Benefit
  - More Floating Point ALU's
  - More routing paths



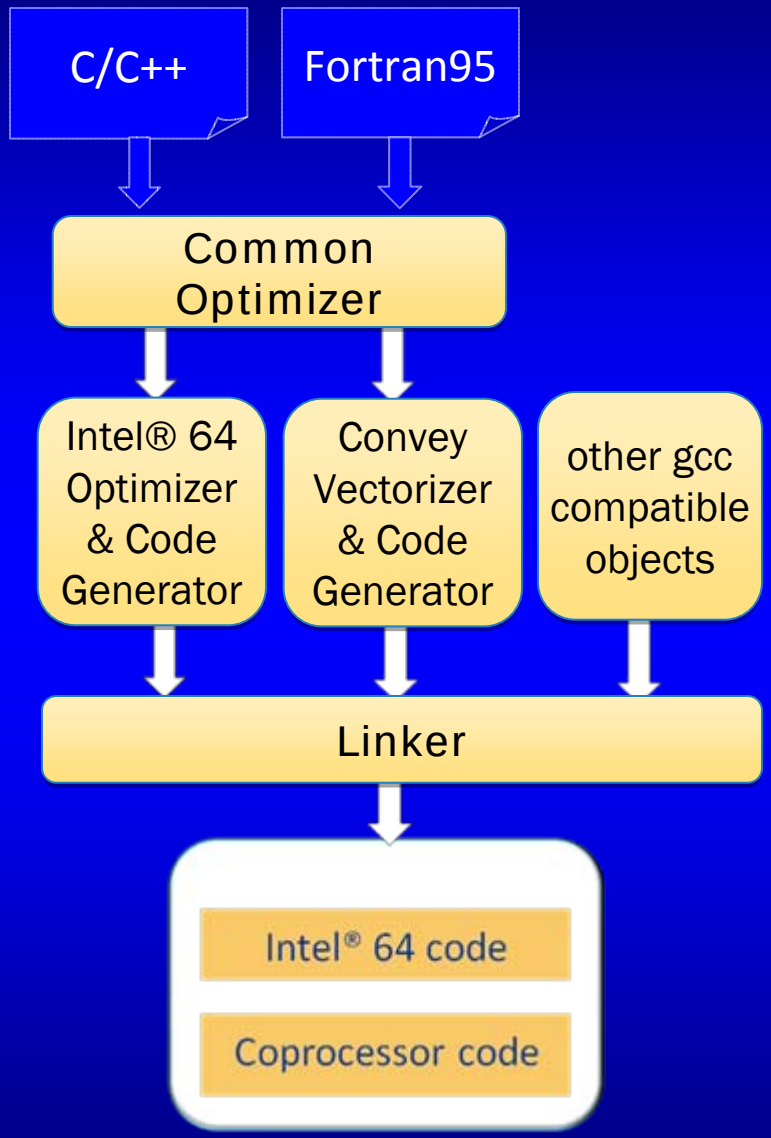
Number of AE FPGA's per node

- Feature Set
  - 4, 8, or 16 as a function of physical memory capacity
    - 4 -One byte per flop – mean factor of 1
    - 8 - ½ byte per flop – mean factor of 2
    - 16 -¼ byte per flop – mean factor of 4
- Benefit
  - More Internal Parallelism
  - Transparent to user
  - Potential heterogeneity within node



Convey Compilers

- Program in ANSI standard C/C++ and Fortran
  - PGAS ready
- Unified compiler generates x86 & coprocessor instructions
- Seamless debugging environment for Intel & coprocessor code
- Executable can run on x86_64 nodes or on Convey Hybrid-Core nodes



Programming Model

Example 4.1-1: Matrix by Vector Multiply

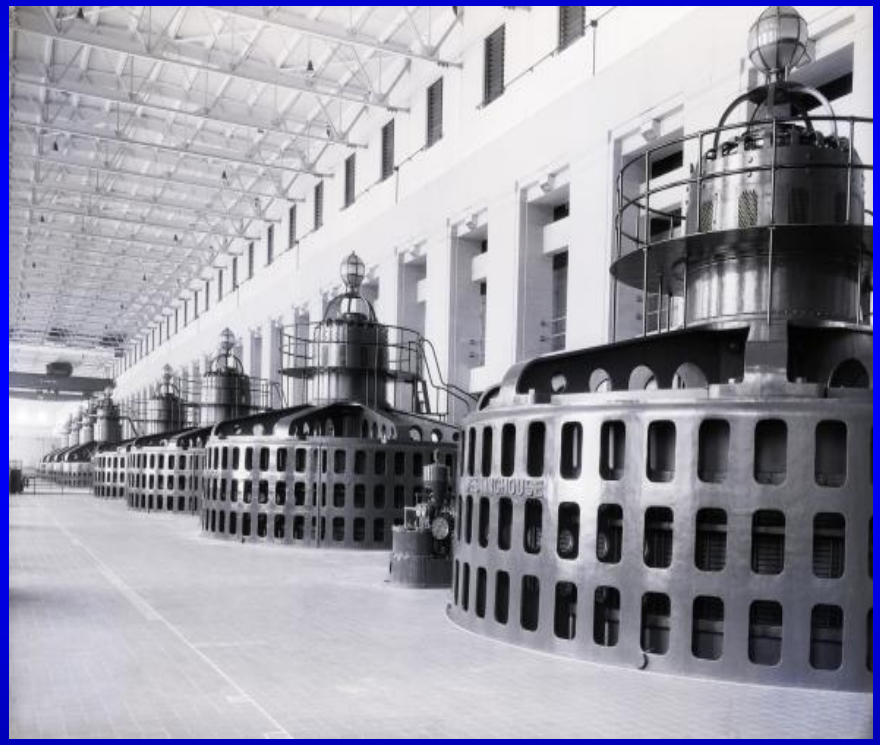
```
#include<upc_relaxed.h>
#define N 200*THREADS
shared [N] double A[N][N];          NOTE: Thread is 16000
shared double b[N], x[N];
void main()
{
int i,j;
/* reading the elements of matrix A and the
vector x and initializing the vector b to zeros
*/
        upc_forall(i=0;i<N;i++;i)
                for(j=0;j<N;j++)
                b[i]+=A[i][j]*x[j] ;
}
```

Math results in

- Using the mean
  - 7 Moore's law
  - 4 Matrix arithmetics
  - .90 efficiency (percentage of peak)
  - 8 Fused multiply/add (64 bits)
  - 4 16_ AE's (user visible pipelines) per node
  - Or a MEAN of 800 times today
    - Best Case – 2304
    - Worst Case - 448

The System

- 2010 base level node is 80 GFlops/node (peak)
- Thus 7 x 4 x .9 x 8 x 4 = 806 Factor
  - Mean of 800 = +1500 (upside)/- 400 (downside)
  - 64 TFlops/node (peak)
  - 16,000 Nodes/Exascale Linpack
- 64 bit virtual address space
  - Flat address space
  - UPC addressing paradigm integrated within TLB hardware
  - Programming model is 16000 shared memory nodes
    - Compiler optimizations (user transparent) deal with local node micro-architecture
- Power is 2 KWatts/Node (3U rack mounted)
  - 32 MegaWatts/system
  - 32 TBytes/Node (288 PetaBytes – system)
  - Physical Memory approx 60% of power

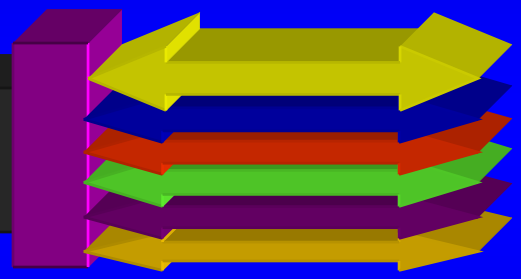


INTEGRATED SMP - WDM

DRAM – 16/32 TeraBytes - HIGHLY INTERLEAVED

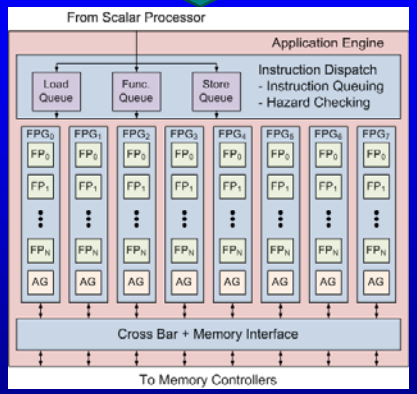
CROSS BAR

MULTI-LAMBDA xmit/receive

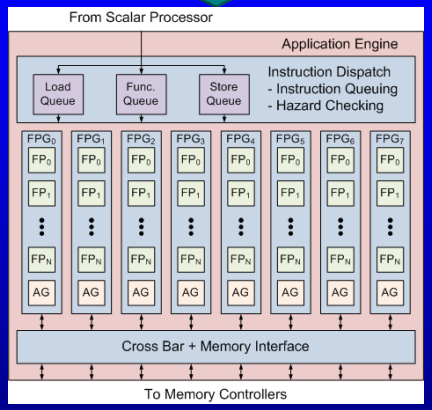


6.4 TBYTES/SEC

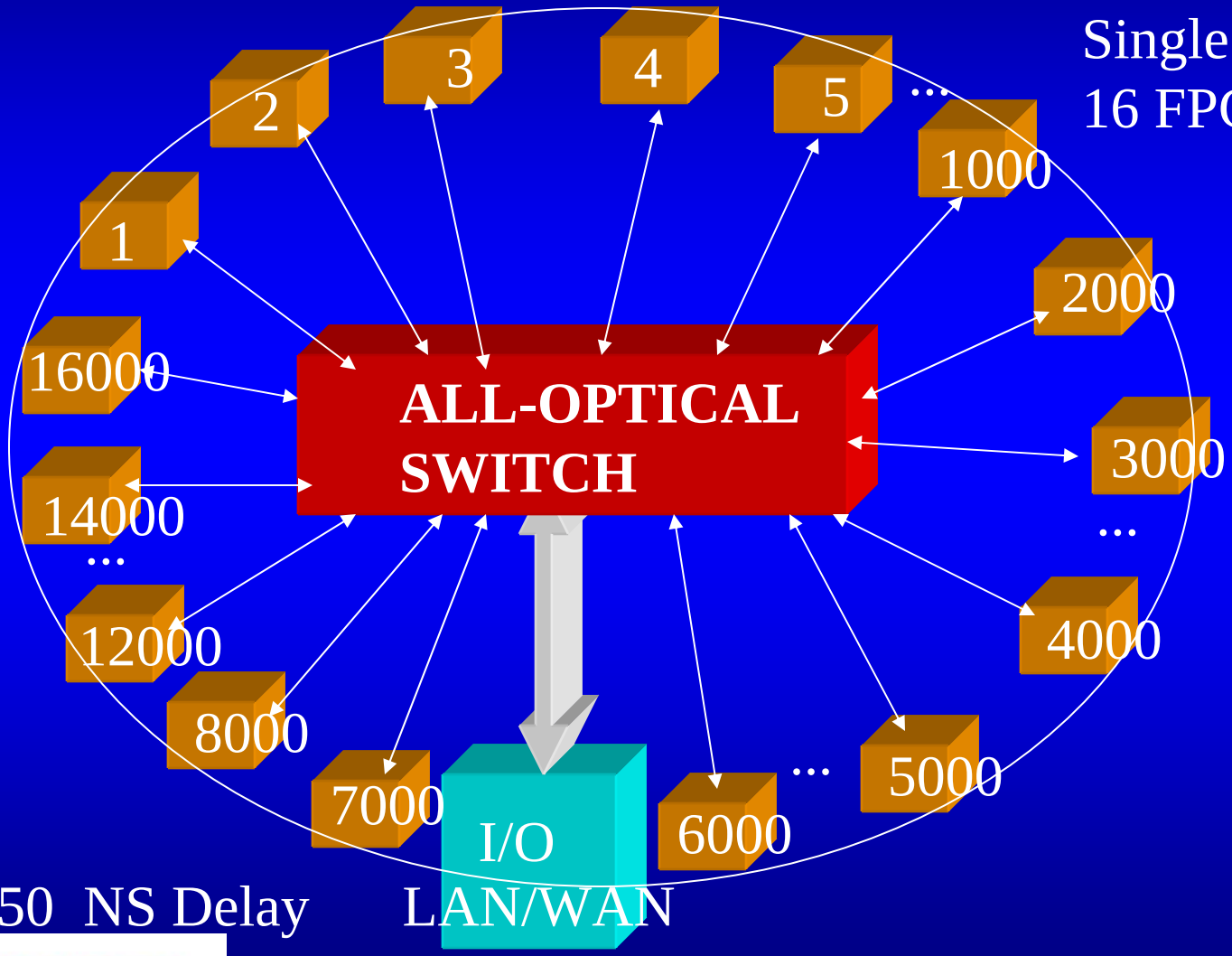
.1 bytes/sec per Peak flop



…



COTS ExaFlop System

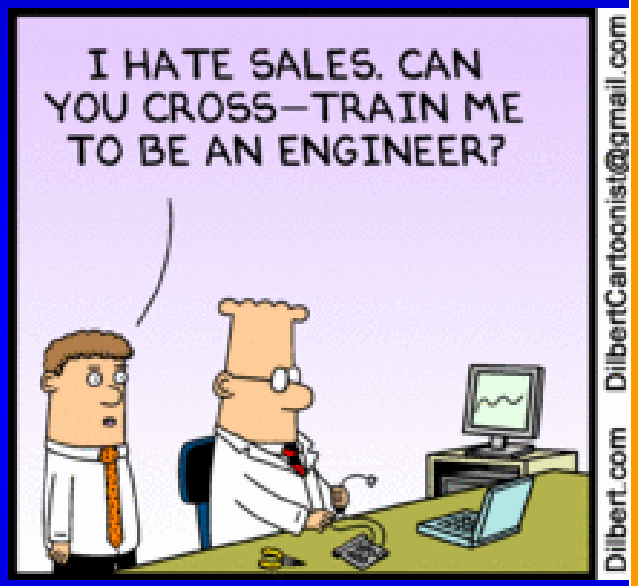


10 meters= 50 NS Delay

What can be done to:

- Reduce power
  - 3D packaging
    - Reduce power in chip to chip interface
    - Better DRAM utilization
    - More power domains (selectively power up/down die regions)
- Architecture (relative to strawman)
  - Less Nodes
    - More matrix arithmetics
      - Twice the performance, same power
    - More floating point IP
      - Twice the performance, same power
  - Less Physical Memory
    - ¼ byte per flops or 15 TBytes/Node yields 20 MWatts
    - (144 PetaBytes Total)

Concluding


I HATE SALES. CAN YOU CROSS-TRAIN ME TO BE AN ENGINEER?
Dilbert.com DilbertCartoonist@gmail.com