K. J. Roche
High Performance Computing Group, Pacific Northwest National Laboratory
Nuclear Theory Group, Department of Physics, University of Washington

- Complexity of the SLDA at a glance

  - FY10 OMB GPRA / PMM benchmark results

- Moving forward

  - identify, sort, and match ...

Bulgac, Drut, Luo, Magierski, Stetcu, Yu
Arbanas, Bertulani, Dean, Kerman

Proudly Operated by Battelle Since 1965

Comment ....

(2011) Tools and Tool Support for the Exascale Era

For the NNSA Workshop on Exascale Computing Technologies (LLNL-TR-472494)

Atinuke Arowojolu, DOE, Sean Blanchard, LANL, James Brandt, SNL, Scott Futral, LLNL, John Mellor-Crummey, Rice University, Barton Miller, University of Wisconsin, David Montoya, LANL Mahesh Rajan, SNLs, Kenneth Roche, PNNL, Martin Schulz, LLNL, Mary Zosel, LLNL

User Needs

- New Debugging Techniques
- Automatic Correlations and Data Analysis in Performance Tools
- Tools for Memory Efficiency and Optimization
- Tools for Threading
- Tools for Power Optimization
- Tools for Transformation to Accelerators

Tool Requirements

- Scalability
- Asynchronous Analysis Capabilities
- Analysis Response
- Fault Tolerance
- Component-based

n := lattice points in one dimension

$$\psi_i(\vec{x}) = \begin{pmatrix} u_i(\vec{x}) \\ v_i(\vec{x}) \end{pmatrix} \rightarrow \begin{pmatrix} u_i(x,y,t)\exp(ik_i z) \\ v_i(x,y,t)\exp(ik_i z) \end{pmatrix}$$

$$m \sim n^3$$

Computation	FP Operations	Data
general solver *	$O(n^9)$	$O(n^6)$
homogeneous solver %	$O(n^6)$	$O(n^6)$
time evolution #	$56mO(mlog_2(m))$	$O(n^6)$

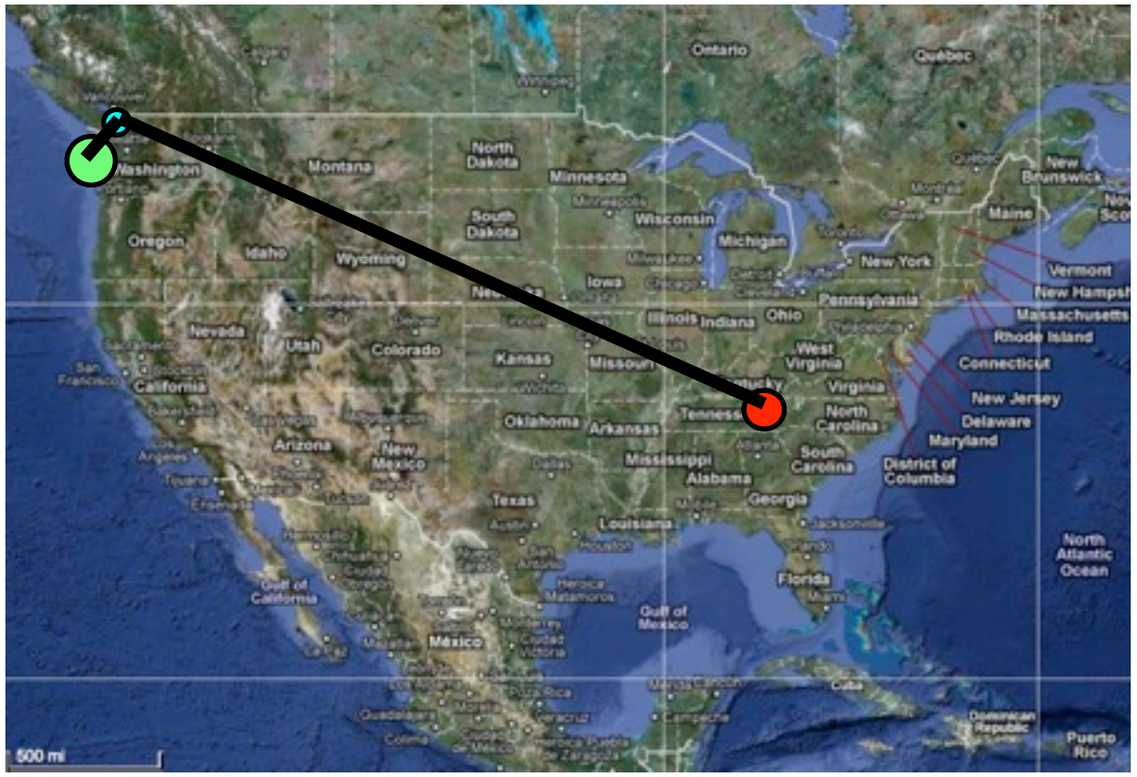
*) per self-consistent iteration; convergence in ~ 10 to 150 iterations
%) solver spatial symmetry to reduce the complexity per iteration; perfect strong scaling
#) per time step ; O(1000) to O(1000000) time steps depending

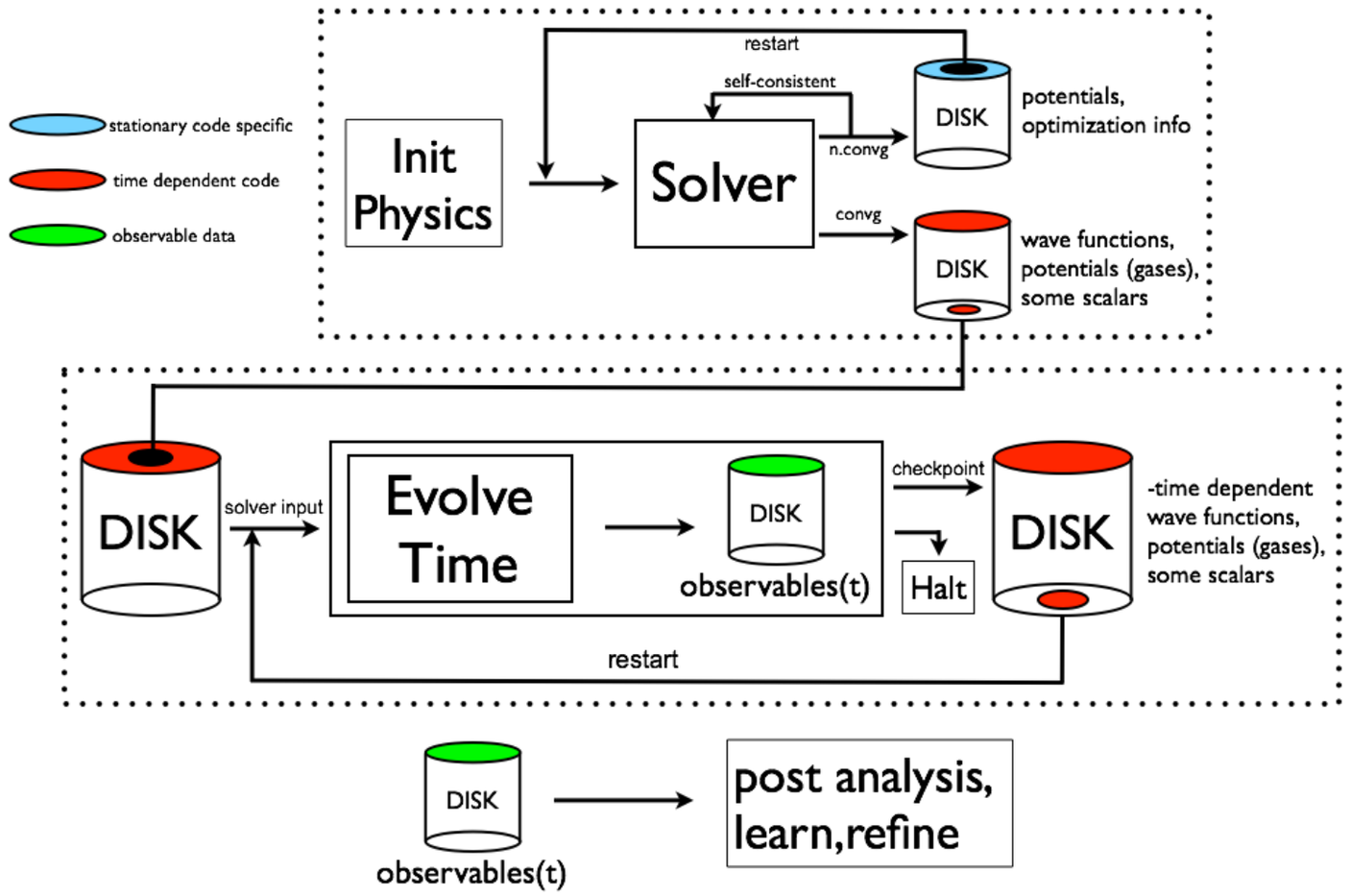
I/O	Data	Other
Observables	$O(n^3)$	per output event; 10 to 100 ts / output; O(100K) ts ; (4,11) observables for (gases,nuclei); double precision type
Checkpoint / restart	$O(n^6)$	at most 1 cp, 1 rs per execution; scaling constant of (22,44) for (gases,nuclei); double precision complex type

--we accumulate all observables for ~ 10 output events in a single file; number of files and overall amount of stored observable data clearly grows w/ number of time steps

--these are usually O(TB)

O() Data Magnitudes: Supercomputers to Laptops





STAGE	k * O() BYTES	TYPE
problem instantiation	**2^10 (k = 1)**	.txt
program text	**2^40 (k = 100)**	binary
ground state wavefunctions	**2^40 (k = 10)**	.bin (.txt, .dat)
checkpointing / progress	**2^40 (k = 10)**	.bin (.dat)
observables	**2^30 (k = 100)**	.txt, .dat, .bin, .silo
movies , plots, etc	**2^20 (k = 10)**	.jpeg, .eps, .m4v

Application	TD-SLDA	POP	LS3DF	Denovo
Problem	***Q2 : Nuclear 198W study*** • Z=74, N=124 • 40 x 40 x 40 lattice • 7,466 p-quasiparticle • 8,946 n-quasiparticle • 200 time steps • 0.75fm spacing • 100MeV cutoff • n = 256000 ***Q4 : Nuclear 238U study*** • Z=92, N=146 • 40 x 40 x 64 lattice • 67,118 p-quasiparticle • 69,508 n-quasiparticle • 200 time steps • 1.25fm spacing • 100MeV cutoff • n = 409600	***3 simulated days, ocean-only model*** • 0.1-degree tripole global grid (3600×2400) • 42 vertical levels • 10 minute time steps • High-frequency output time slice	***Self-consistent DFT calculation for ZnO nanorod*** • 2776 atoms • 24220 valence electrons, d-electrons in valence band • 720×300×300 numerical grid	***Q2 : Full Core EDF PWR900 benchmark*** • 17x17 fuel assemblies • 17x17 fuel pins per assembly • 2x2 cells per pin cell • 3 fuel enrichments • 45 homogenized pin cell materials per assembly • 135 different pin cell materials • 233,858,800 (578x578x700) cells • 168 angles, 1 moment, 2 energy (fast and thermal) groups • 7.86×10^{10} total unknowns ***Q4 : Full Core EDF PWR900 benchmark*** • 168 angles, 1 moment, 44 energy (fast and thermal) groups • 1.73×10^{12} total unknowns
Hardware (cores)				
Q2	(s)73,728; (td)16,414	4,800	43,200	17,424
Q4	(s)217,800; (td)136,628	9600	86,400	112,200
Time (seconds)				
Q2	(s)6538.5, (td)2084.4	957.8	13,932	11,260.8
Q4	(s)18393.2, (td)2031.5	290.3	5328	1121.6
Metric target	(s)Q2:Q4 efficiency ≥ 1.0 ; (td)Q2:Q4 time ≥ 1.0	Q2:Q4 time ≥ 2.0	Q2:Q4 time ≥ 2.0	Q2:Q4 efficiency ≥ 1.0
Metric result	(s)Q2:Q4 efficiency = 2.11 (td)Q2:Q4 time = 1.026	Q2:Q4 time = 3.2992	Q2:Q4 time = 2.6	Q2:Q4 efficiency = 31

Targeted Computing Platforms

Hex-Core AMD Opteron (TM)	2.6e9 Hz clock	4 FP_OPs / cycle / core 128 bit registers
PEs	18,688 nodes	224,256 cpu-cores (processors) [+74752]*
Memory	16 GB / node 6 MB shared L3 / chip 512 KB L2 / core 64 KB D,I L1 / core	dual socket nodes 800 MHz DDR2 DIMM 25.6 GBps / node memory bw
Network	AMD HT SeaStar2+	3D torus topology 6 switch ports / SeaStar2+ chip 9.6 GBps interconnect bw / port 3.2GBps injection bw
Operating Systems	Cray Linux Environment (CLE) (xt-os2.2.41A)	SuSE Linux on service / io nodes

FY	Aggregrated Cycles	Aggregated Memory	Aggregated FLOPs	Memory/FLOPs
2008	65.7888 THz	61.1875 TB	263.155 TF	0.2556
2009	343.8592 THz	321.057 TB	1.375 PF	0.2567
2010 / 2011	583.0656 THz	321.057 TB	2.332 PF	0.1513

* pre-Titan upgrade prior to end of FY11 (16 core + Gemini + ...)

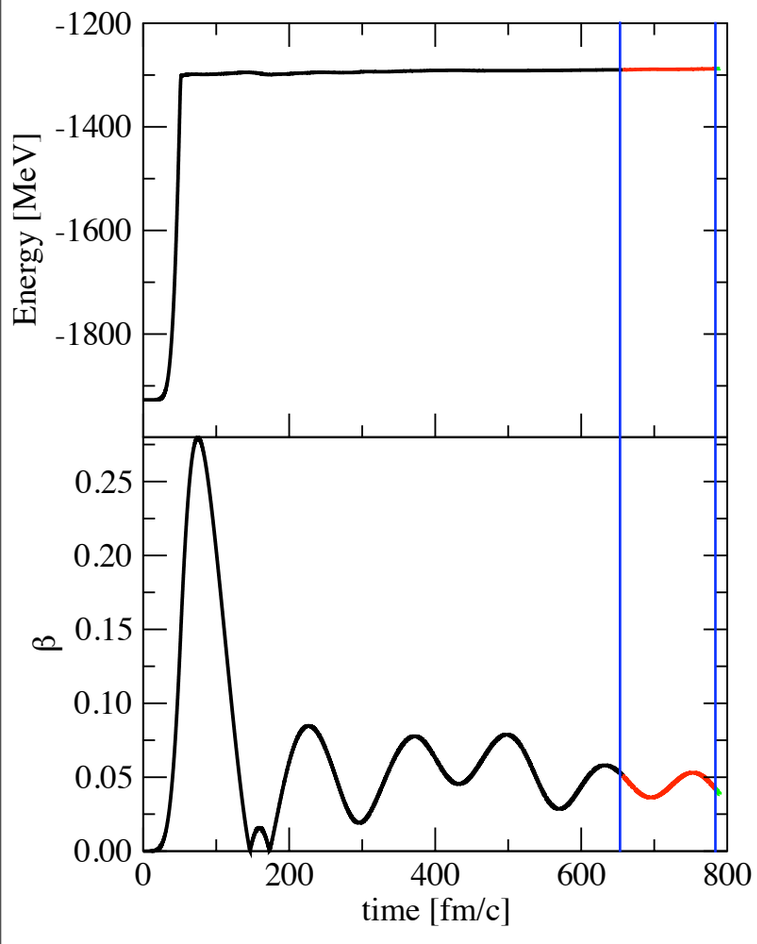
Nuclear TD-SLDA

E1:

Machine Event (16,414PEs)	Phase 1	Phase 2
Instructions Retired	76088401412884728	488719911422539 43
Floating Point Ops	2020709113712	1601488874412091
Time (s)	1125.904417	958.519619

E2:

Machine Event (136628PEs)	Phase 1	Phase 2
Instructions Retired	1.7285585098 40934e+17	537846559382408604
Floating Point Ops	147028670495	18891863365740396
Time (s)	362.958407	1386.519066



Scaling Behavior:

-scaled problem -same Skyrme functional (SLy4)
- **-- 198W , 238U**
- -- 40^3 , 40^2 x 64 := 1.6x more complex in spatial data
- -- 16412 wfs, 136628 wfs

- 200 time steps
- --predicted Q4 work rate := (time / ts (Q2)) * 1.6 ~ 7.668156952 s/ts;
- measured rate 6.93259533 s/ts 110.61%

```
PEs              = 8.323869867186548 ( 136628 / 16414 )
Time             = 0.974629425161743 (2031.541/ 2084.424036)
INS              = 5.687421396767019 (7.10702410366502e+17/ 1.249603925551387e+17)
FP               = 11.781663538842758 (1.889201039441089e+16/1603509583525803)
QPWFs            = 8.324762368998294 ( 136626 / 16412 )
CMPLX / WF       = 1.6 (4x40x40x64 / 4x40x40x40)
Time IO(rd)      = 0.322370532986372 (362.958407/1125.904417)
BYTES IO(rd)     = 13.31961979039727 (4x40x40x64x136626x16 / 4x40x40x40x16412x16)
```

E2:

Machine Event (136628PEs)	Phase 1	Phase 2
Instructions Retired	1.728558509840934e+17	537846559382408604
Floating Point Ops	147028670495	18891863365740396
Time (s)	362.958407	1386.519066

Raw Data:

```
real 33m51.541s
user 0m27.878s
sys 0m1.740s

         Time[ms]         INS                  FP                 DCM
Init:    32228731   9045196589846562          91057643393        363095113632
I/O:    330729676  163810654394246847          55971027102       1848702550452
T_loop: 1386519066  537846559382408604   18891863365740396   528423132128308
```

2031.541 s - 1749.477 s= 282.064 s ??

Costs

DATA \ CODE	SOLVER	TD (200ts)	Total
INS	1.37E+19	7.11E+17	1.44E+19
FP_OP	9.42E+17	1.89E+16	9.61E+17
Wall time	18393.181	2031.541	20424.722
$ CPU Hrs	1,112,787	77,101	1,189,888
PEs	217800	136628	

Time Stepping Only ...

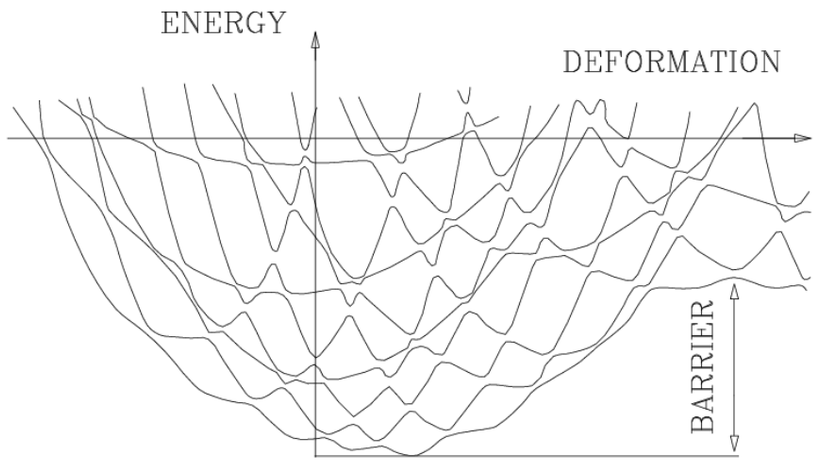
18891863365740396 FP_OPs / (200 TS * 6.93259533 S / TS) ~ 13,625,390,900,334 <FLOPs>

Advanced Computing Techniques + Exascale

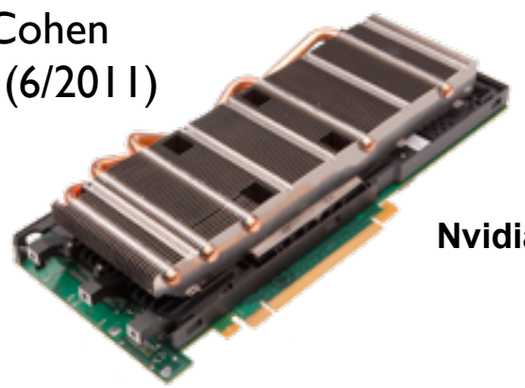
Extreme Complexity

FNCs :	**718,112,000**
PEs :	**102,590,000**
MEM :	**1e+17 to 1e+18 B**
INS / TS :	**5.3 e+20**
FP_OP / TS :	**2.0 e+20**

- adjust to new hardware designs and scales -hybrid approaches
- (stochastic) SLDA : OpenCL, FUSION, (m-core/GPU) --is it necessary
- improve lattice boundary conditions, iterative / higher and mixed precision numerics
- simple, novel data structures for complex memory hierarchies (ht)
- specialized collective communications on the lattice
- fault detection and mitigation
- advanced CPRS and DATA analysis techniques / content extraction from enormous numbers of extremely large 'files'
- novel end-end data movement and reduction techniques from supercomputer to laptop

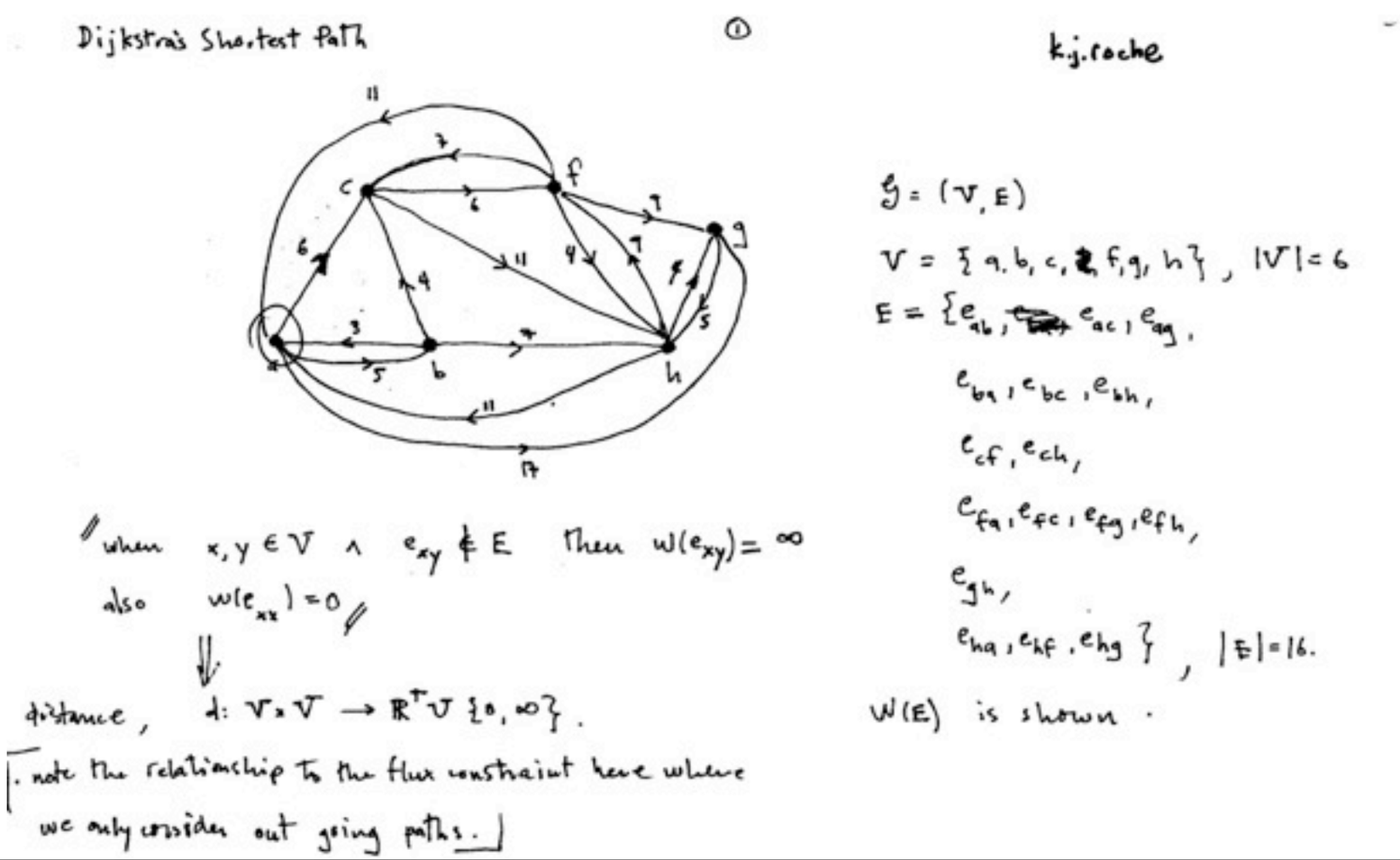


w/ Saul Cohen
UW PD (6/2011)



Nvidia M2050 GPU

1/20th the power consumption and 1/10th the cost



- Effective at scale use of the biggest US open science supercomputer to achieve new science

  - first implementation of the parallel td-slda in 3d for both ufg and nuclear systems; prototyped the parallel homogeneous solver and parallel nuclear solver
    -ALL newer versions are descendants of these codes

  - fully scalable check point and restart capability (~ 3 GBPS small and 20 GBPS large events)
    - LUSTRE > POSIX > MPI

  - pipelined data analysis from human thought to supercomputer to movies of evolving systems

  - prototyped a direct parallel complex symmetric diagonalization routine for the KKM

  - recently prototyped some parallel stochastic evolution codes for real and imaginary times
    - designed a 'bit'-based data structure that allows dynamic allocation and element access w/ overhead of at most `unsigned char` storage demands
    - Boolean logic based operations such as *toggle by index*

- PI of DOE SC ALCC award on JaguarPF (150M CPU-Hrs in FY10, 100M CPU-Hrs in FY11)

- 7/2007-8/2007, invited speaker, CScADS (DOE Center for Scalable Application Development Software)
- J.Phys. Conf. Ser. 012064 (2008)
- SciDAC 2008, invited poster
- arxiv.org/abs/1011.5999
- 2010, PNNL ALD Invited Poster, Dan Hitchcock, DOE ASCR AD
- Science, 10 June 2011:Vol. 332 no. 6035 pp. 1288-1291 DOI: 10.1126/science.1201968
- SciDAC 2011, JPCS submission (Stoitsov et al)
- CPC in preparation