CS Research Nuggets for Exascale Data Analytics

DOE Workshop on
Exascale Data Management, Analysis & Visualization
Houston, TX
February 22-23, 2011

Nagiza F. Samatova
samatovan@ornl.gov
Oak Ridge National Laboratory
North Carolina State University

February 23, 2011

Acknowledgements

- **DOE ASCR for funding CS research in data analytics**
- **NSF Expedition in Computing**
- **Arie Shoshani, Alok Choudhary, Rob Ross, Scott Klasky, Vipin Kumar, PhD students, and more**
  - PI's and co-PI's on the projects
- **LCF Facilities at ORNL**
- **Application Scientists:**
  - CS Chang, PPPL
  - Stephane, PPPL
  - Fred Semazzi, NCSU
  - Many others

WHAT IS DATA ANALYTICS AT EXASCALE?

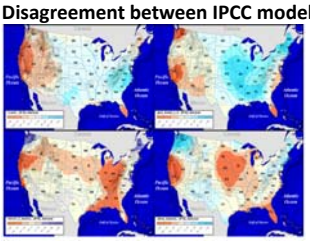
- **Scales to 200K cores?**
- **Processes PBs data in a reasonable time?**
- **As simulation proceeds, reduces PBs data by orders of magnitude but ensures information quality?**
- **In situ extracts features so that post-processing analysis could be performed in O(N) vs. O(N^2)?**
- **Transforms a complex problem into a form that makes many analysis techniques practical?**
- **…**

PROVIDES DATA-DRIVEN UNDERSTANDING OF GRAND-CHALLENGE SCI. APPLICATION PROBLEMS

PHYSICS-BASED MODELS ARE ESSENTIAL BUT NOT ADEQUATE

- **Models make relatively reliable predictions at global scale for ancillary variables:**
  - Sea Surface Temperature (SST)
  - Temperature/humidity profiles over land
  - Wind speed at different heights
- **They provide *least* reliable predictions for variables that are crucial for impact assessment:**
  - Regional precipitation and extremes
  - **Hurricane intensity and frequency**
  - Droughts and floods

Disagreement between IPCC models



Regional hydrology ("P–E" changes in 2030s) exhibits large variations among major IPCC model projections

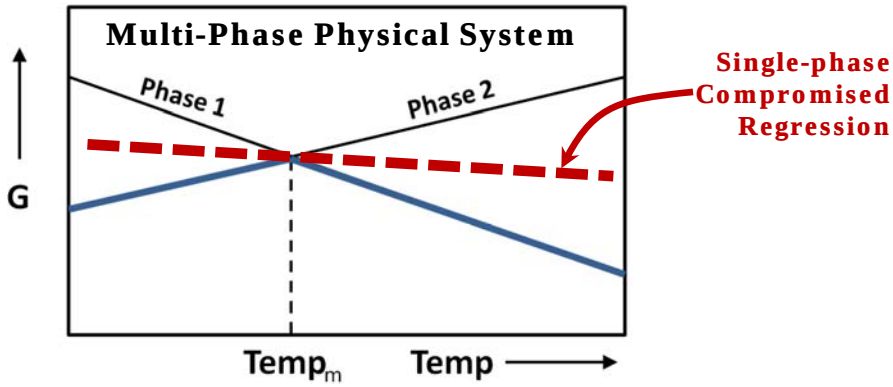
"The sad truth of climate science is that the most crucial information is the least reliable" (Nature, 2010)

Hypothesis-driven "manual" conceptual models try to address this gap:
- **Hurricane models (Emanuel et al, BAMS, 2008)**
- **Regional-scale precipitation extremes (O'Gorman & Schneider, PNAS, 2008; Sugiyama et al, PNAS, 2010)**

We need a systematic approach to hybrid data-driven + physics-based inference.

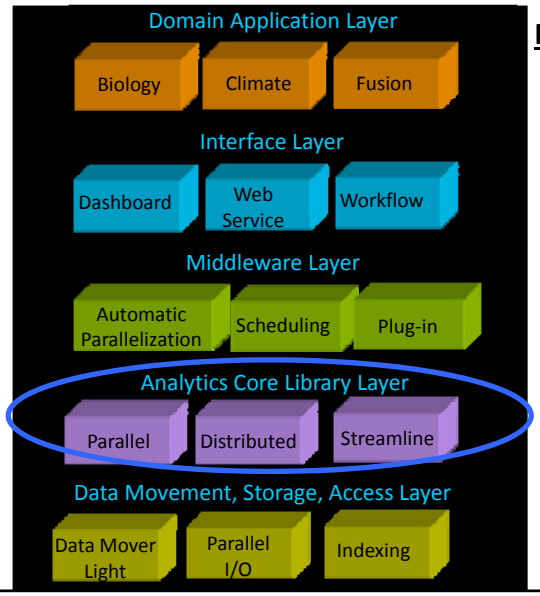
FAMILY = DOMAIN SCIENTISTS + CS RESEARCHERS

Hurricane activity is the climate system's response initiated by a liquid-vapor phase transition associated with non-linearly coupled fluctuations in the ocean and the atmosphere.



- Gibbs free energy (1876)
- Landau's theory of phase transitions (1936)

END-TO-END DATA ANALYTICS SOFTWARE STACK IS COMPLEX: GENERIC (ALL APPLICATIONS) PERSPECTIVE



Exemplars

CDAT

Kepler, VisIt, SciRun, FIESTA

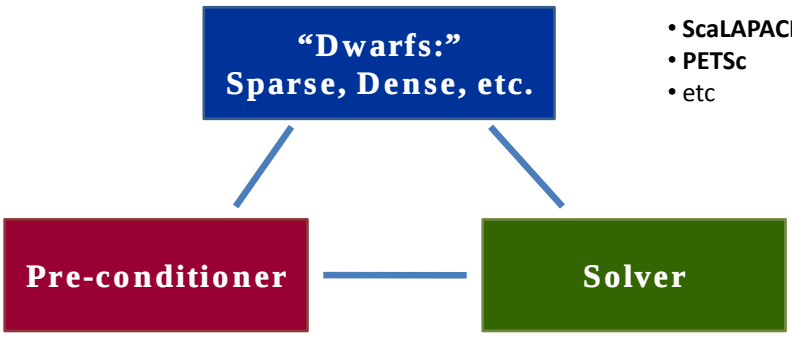
pR, VisIt

RScaLAPACK, NW-Minebench

FastBit, ADIOS pnetCDF, HDF5

Focus of my talk

THE LESSON LEARNED FROM LINEAR ALGEBRA



- ScaLAPACK
- PETSc
- etc

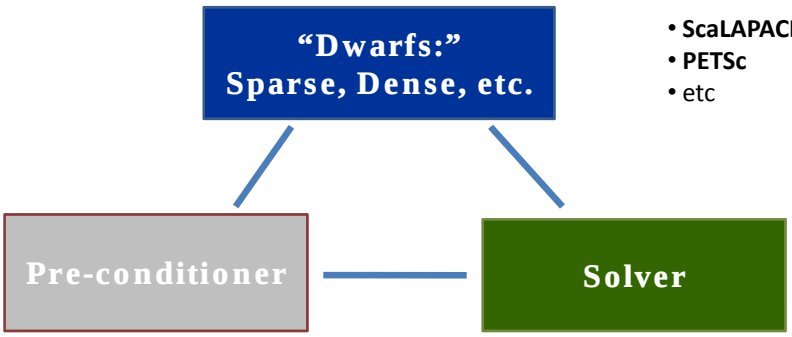
"...conditions a given problem into a form that is more suitable for numerical solution."

highly "optimized" computational kernel

SOFTWARE FOR DATA ANALYTICS IS MORE AD HOC

- **Should we adopt this approach from Linear Algebra to Data Analytics at extreme scale? If so, then**
  - What are the "Dwarfs" for data analytics?
  - What about the "Preconditioners?"
  - What are the "Computational kernels?"
- **Do we/should we have a ScaLAPACK-like library for Exascale Data Analytics?**

THE LESSON LEARNED FROM LINEAR ALGEBRA



- ScaLAPACK
- PETSc
- etc

"...conditions a given problem into a form that is more suitable for numerical solution."

highly "optimized" computational kernel

"COMPUTATIONAL KERNELS" CONCEPT IS PROMISING

The frequency of kernel operations in illustrative data mining algorithms and applications.

Application	Top 3 Kernels (%)			Sum %
	Kernel 1 (%)	Kernel 2 (%)	Kernel 3 (%)	
K-means	Distance (68)	Center (21)	minDist (10)	99
Fuzzy K-means	Center (58)	Distance (39)	fuzzySum (1)	98
BIRCH	Distance (54)	Variance (22)	redist.(10)	86
HOP	Density (39)	Search (30)	Gather (23)	92
Naïve Bayesian	probCal (49)	Variance (38)	dataRead (10)	97
ScalParC	Classify (37)	giniCalc (36)	Compare (24)	97
Apriori	Subset (58)	dataRead (14)	Increment (8)	80
Eclat	Intersect (39)	addClass (23)	invertC (10)	72
SVMlight	quotMatrix(57)	quadGrad (38)	quotUpdate(2)	97

Alok Choudhary, NWU, *NU-Minebench*

IF WE ARE LUCKY…

- **and Jack Dongara did most of the work for us:**
  - Some data analysis routines call linear algebra functions
  - In R, they are built on top of LAPACK library
- **RScaLAPACK is an R wrapper library to ScaLAPACK**

```
A = matrix(rnorm(256),16,16)
b = as.vector(rnorm(16))
```

Using RScaLAPACK:

```
library (RScaLAPACK)
sla.solve (A,b)
sla.svd (A)
sla.prcomp (A)
```

Using R:

```
solve (A,b)
La.svd (A)
prcomp (A)
```

WHAT IF WE ARE NOT THAT LUCKY?

KERNEL "OPTIMIZATION METRICS" GET COMPLEX

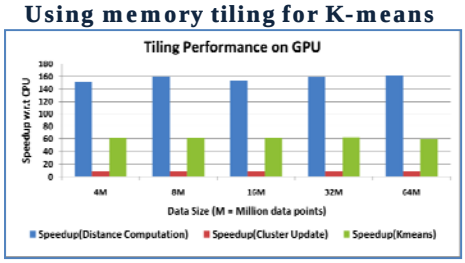
- **Depend on target data processing regimes**
  - In-situ, In-staging, Post processing, Out-of-core, Multi-resolution
  - Data streams, Distributed
- **Depend on target performance criteria**
  - Strong/Weak Scalability, Memory & Data Movement Requirements
  - Energy Efficiency, Resiliency, Concurrency, Fault Tolerance
  - Numerical stability, Parameter-free
- **Depend on target layer in SW stack**
  - Simulation Code, Parallel I/O Library, Active Storage
  - VisIt, Parallel R, Matlab, Python, stand-alone
- **Depend on target HW architecture**
  - Hybrid multi-node, multi-core HPC architectures comprised of a mix of GPUs, FPGA, and SSDs
  - Clouds

KERNEL "OPTIMIZATION METRICS" GET COMPLEX

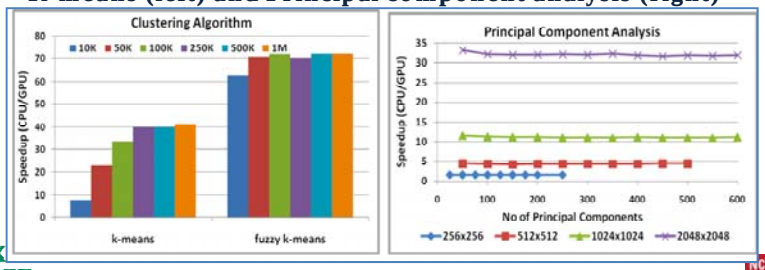
- **Depend on target data processing regimes**
  - In situ, In-staging, **Post processing**, Out-of-core, Multi-resolution
  - Data streams, Distributed
- **Depend on target performance criteria**
  - Strong/Weak **Scalability**, **Memory & Data Movem't** Requirements
  - Energy Efficiency, Resiliency, Concurrency, Fault Tolerance
  - Numerical stability, Parameter-free
- **Depend on target layer in SW stack**
  - Simulation Code, Parallel I/O Library, Active Storage
  - VisIt, Parallel R, Matlab, Python, **stand-alone**
- **Depend on target HW architecture**
  - **Hybrid multi-node, multi-core HPC architectures** comprised of a mix of **GPUs**, FPGA, and SSDs
  - Clouds

PERFORMANCE IMPROVEMENTS ON GPUS OVER CPUS

Using memory tiling for K-means

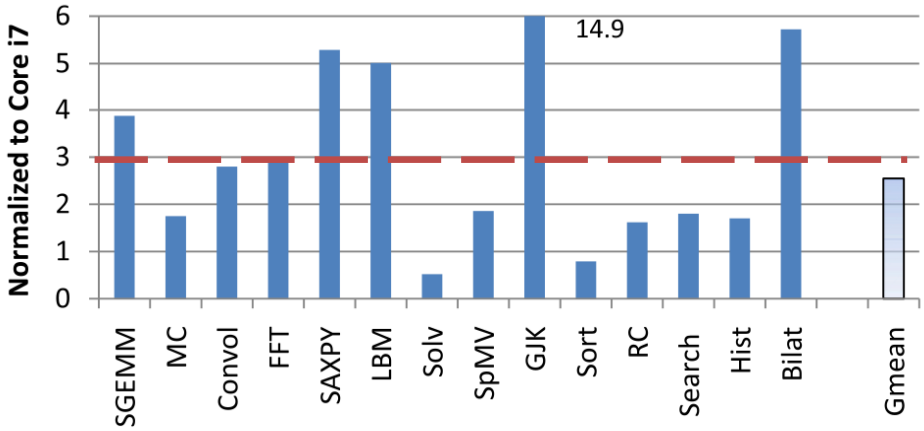


K-means (left) and Principal component analysis (right)



Alok Choudhary, NWU, *NU-Minebench*

PREVIOUS SUCCESSES (AND FAILURES) ON GPU



V. W. Lee, C. Kim et al. Debunking the 100X GPU vs. CPU myth: An evaluation of throughput computing on CPU and GPU. *Int'l Symposium on Computer Architecture*, pages 451-460, 2010

COMPUTATIONAL MOTIFS

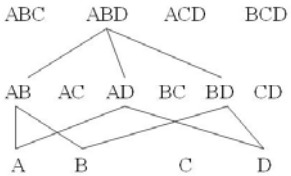
- Dense Linear Algebra
- Sparse Linear Algebra
- Spectral Methods
- N-Body Methods
- Structured Grids
- Unstructured Grids
- MapReduce
- Combinational Logic
- Graph Traversal
- Dynamic Programming
- Backtrack and Branch-and-Bound
- Graphical Models
- Finite State Machines

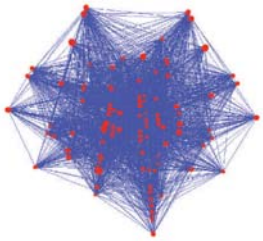
Accelerated Weakly-Accelerated Hard-to-Accelerate

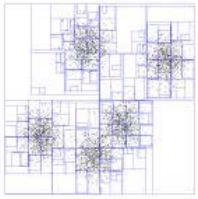
J. D. Owens. What's New with the GPU? GPUs for Scientific Computing and Visualization. *Presented at SciDAC 2010*

Backtracking Ubiquity in Data Mining

- Game-playing
- Constraint Satisfaction
- Frequent Itemset Mining
- k-d Tree Construction, Traversal
- Graph coloring/partitioning



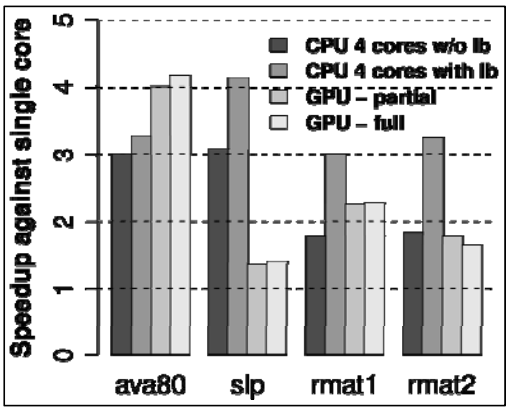




Challenges: Backtracking ↔ GPU characteristics

	Backtracking (worst-case)	GPU Optimal
Problem Instance	Irregular access pattern (e.g. sparse matrix mult)	regular access with locality (e.g. dense matrix mult)
Work Unit	variable in size and computation (e.g CSP on large sets)	constant size, SIMD (e.g. stream processing)
Output	exponential size, hard to estimate (e.g. subset enumeration)	polynomial size, apriori (e.g. dense matrix mult)
Search Space	Tree-based, unbalanced (e.g. 8-queens)	Fixed, apriori (if applicable) (e.g. k-d trees)

BACKTRACKING: 1.4--2.25 TIMES A SINGLE CPU CORE



- Combinational Logic
- Graph Traversal
- Dynamic Programming
- Backtrack and Branch-and-Bound
- Graphical Models
- Finite State Machines

Accelerated **Weakly-Accelerated** **Hard-to-Accelerate**

KERNEL "OPTIMIZATION METRICS" GET COMPLEX

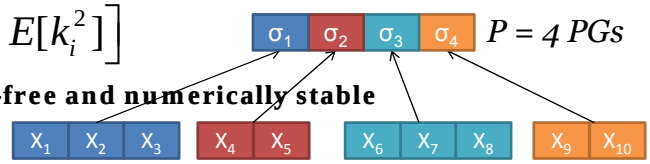
- **Depend on target data processing regimes**
  - **In situ**, In-staging, Post processing, Out-of-core, Multi-resolution
  - **Data streams**, Distributed
- **Depend on target performance criteria**
  - Strong/Weak Scalability, Memory & **Data Movem't** Requirements
  - Energy Efficiency, Resiliency, Concurrency, Fault Tolerance
  - **Numerical stability**, Parameter-free
- **Depend on target layer in SW stack**
  - Simulation Code, **Parallel I/O Library (ADIOS)**, Active Storage
  - VisIt, Parallel R, Matlab, Python, stand-alone
- **Depend on target HW architecture**
  - **Hybrid multi-node, multi-core HPC architectures** comprised of a mix of GPUs, FPGA, and SSDs
  - Clouds

In-Situ Communication-free, Numerically Stable Variance

1. $\sigma^2 = E(X - E(X))^2$
   Numerically stable, but $X - E(X)$ requires **communication**

2. $\sigma^2 = E(X)^2 - E(X^2)$:
   Communication-free, but numerically unstable

3. $$\sigma^2 = \sum_{i=1}^{P}\left[\sigma_i^2 + E[k_i^2]\right]$$



   Communication-free and numerically stable

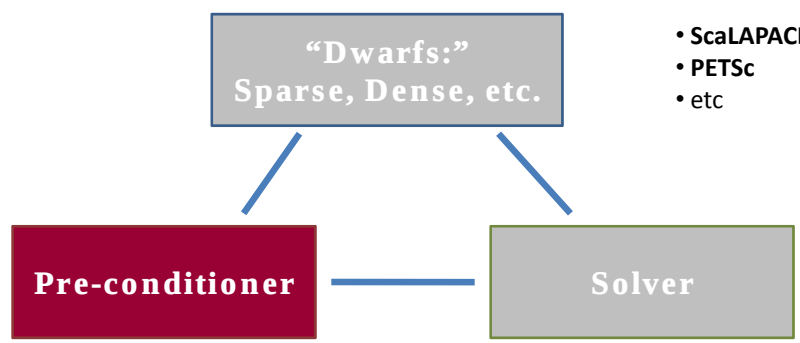
   Each Processor Group i saves local variance σ_i^2 and local mean m_i, where global mean $G = L_i + k_i$.

   $E(X - G)^2 = E(X - (L_i + k_i))^2$

   $= E[(X - L_i)^2] + E[k_i^2] - \boxed{2 * k_i * E[X - L_i]} \rightarrow 0$

   $E[(X - L_i)^2]$, numerically stable + communication-free, calculated at write time; $E[k_i^2]$ is calculated at read time.

The Lesson Learned from Linear Algebra



- ScaLAPACK
- PETSc
- etc

"…conditions a given problem into a form that is more suitable for numerical solution."

highly "optimized" computational kernel

"PRECONDITIONERS" FOR DATA ANALYTICS

- **How to define a "preconditioner" for data analytics?**

Solve a Problem P_{hard}

Directly

Indirectly (via "Preconditioner"):

Reduce a Hard Problem P_{hard} to a "Better" Problem P_{better}

$$P_{hard} \rightarrow \text{Preconditioner} \rightarrow P_{better}$$

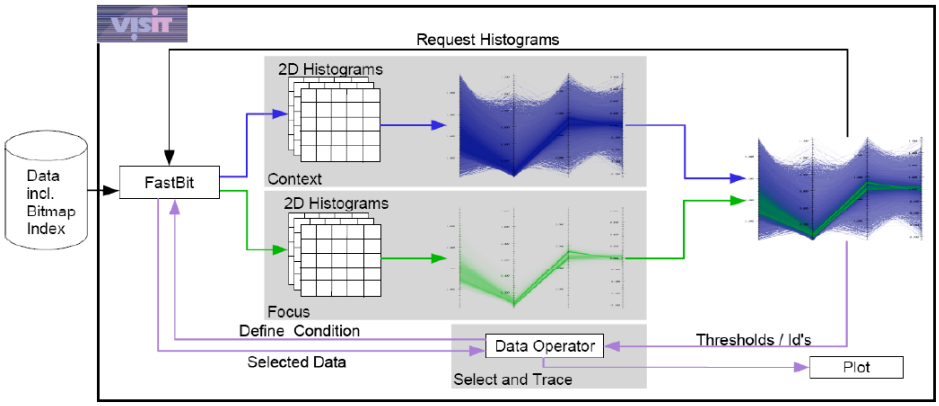
"Better" in terms of:

- Increased throughput
- Faster time-to-solution
- More accurate solution
- Higher data compression rate
- Approximate but real-time solution

23

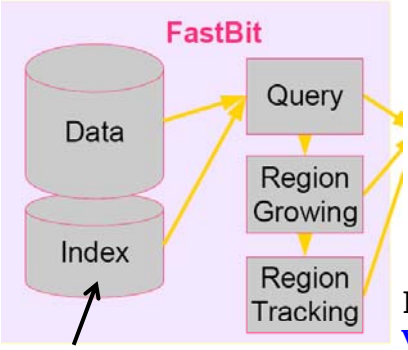
INDEXING AS A PRECONDITIONER FOR INDEX-BASED ANALYTICS

INDEX-BASED ANALYSIS IN VISIT



[Rübel, Prabhat, Wu, Childs, Meredith, Geddes, Cormier-Michel, Ahern, Weber, Messmer, Hagen, Hamann and Bethel. SC08]

CHALLENGES FOR INDEX-BASED ANALYTICS



0.3—1.5X of Data

Multiple Variables over Different Spatial Resolution

Index over Common Spatial Resolution

Query

Regions
Very Fast
Specific Analysis (e.g., histogram)

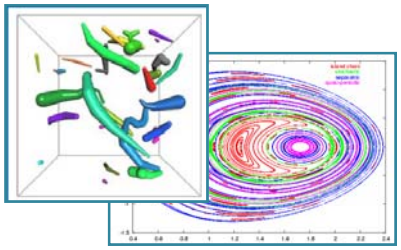
Quantized Data
Which analyses are robust to quantization?

Raw Data
I/O-bound

THINGS GET WORSE AS INDEXING DERIVED OBJECTS

Raw Data

Features Extracted



- Deborah Silver, Rutgers
- Valerio Pascucci, U. Utah
- Chandrika Kamath, LLNL

Compressed/Transformed Data (Multi-Res.)

Different Interpolations/Extrapolations

Processed by Different Analyses and Parameters

Feature & Event classification for Fusion-

Feature Based Techniques to characterize and catalogue interesting phenomena

Objects/Features	Move	Interact ---Come Together/Apart
(Plasma specific) Blobs Filaments Avaloids Striations Bursts Radial streamers IPO (Intermittent plasma Objects) Holes (opposite of blobs, density rarefactions) Chaotic Field line regions Wiggles Flux tubes Loss Cone *(CFD-general)* bubble, hole, blast wave, packet, blobcloud, patch, critical pt., point, eddy, ring, Favor, roll, filament, separatrix, finger, spike, gyres, spiral, hairpin, striation, helix, vortex	*(Plasma specific)* Spin/wake (blobs) Zonal flows Flow shears Rotation *(CFD general)* advect swirl entangle transport disperse, wind flow hop migrate stream	*(Plasma specific)* Coalesce (blobs) Breakup (blobs) *(CFD General)* accrete, aggregate, align, bind, bifurcate, burst, collapse, condense, disassemble, disrupt, finger, fission, focus, fold, fuse, pair, roll-up, plow, reflect, scatter, spike, split, striate, strip, wind about

Deborah Silver, Rutgers

In Situ Preconditioners for Scientific Data Compression

- **Myth: "Scientific data is almost uncompressible."**

GTS Fusion Simulation Data (Stephane, PPPL)

C&R Data
- ~2TB per C&R
- Every 1 hour
- Two copies
- Keep the last copy

Analysis Data
- ~2TB per run (now)
- Every 10th time step
- Cannot afford storing all b/s of
- Analysis routines and I/O reads
- Matlab analysis routines

V&V Data
- Small
- Every 2nd time step

Expected: 10-fold increase by 2012-2014

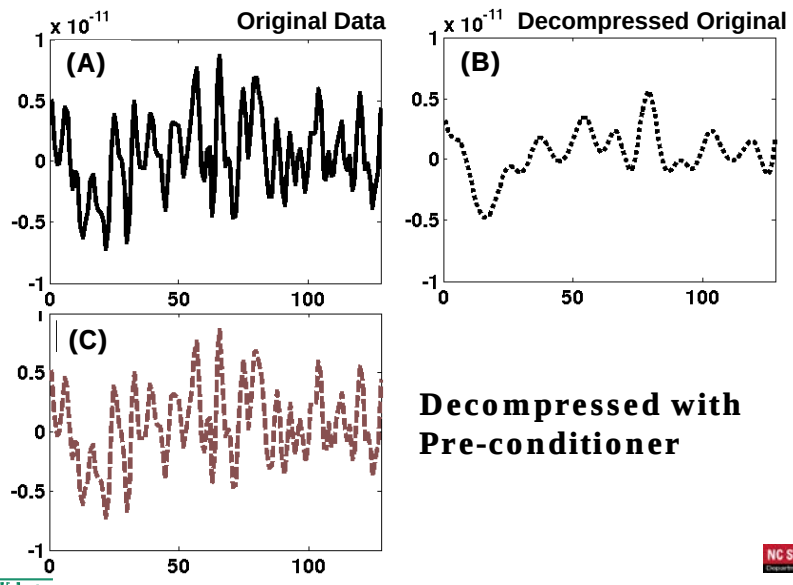
Computing and Storage Resources

	GYRO		GTS		XGC1*	
Facilities	NERSC/OLCF		NERSC/OLCF		NERSC/OLCF	
Architectures	XT5,Power,Cluster		XT5		XT5	
Years	Present	In 5 yrs	Present	In 5 yrs	Present	In 5 yrs
Hrs used/year	30M	50M	24M	50M	65M	500M
NERSC'09 used	1.2Mhrs		~2Mhrs		~8M hrs	
#Cores per run	512	512	8-98K	32-130K	10-223K	1M
Wall clock/run	12	24	72 Hrs	72 Hrs	20-100hrs	20-100hrs
Memory/run	512GB	1.024TB	16-100T	32-160TB	40 TB	100 TB
Min Memory/core	1GB	2GB	1GB	1GB	0.3GB	0.1GB
Read/Write data			2.5TB	8TB	5TB	25TB
Checkpoint size	4GB	8GB	1-8GB	1-10 GB	1TB	5TB
Data in/out nersc			5GB/run	10GB/run	10GB/day	50GB/day
On-line storage			4TB/10K	8TB/10K	4TB/3K	5TB/3K
Off-line storage			25GB	100GB	1TB/30	10TB/100

FROM C.S. CHANG'S TALK AT NERSC

*Unstructured mesh

EFFECT OF PRECONDITIONER ON WAVELET COMPRESSION



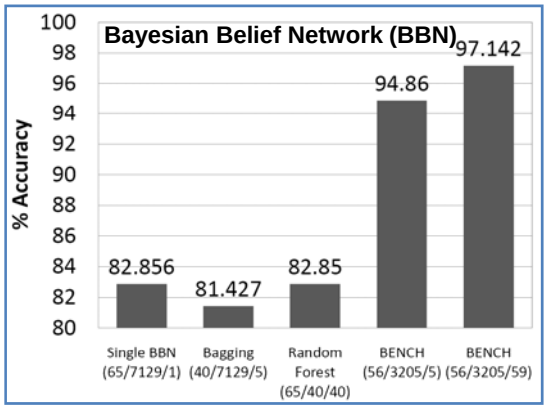
Decompressed with Pre-conditioner

ANALYSIS DATA COMPRESSION W/ ISABELA

- **Analysis data is stored every N-*th* time step:**
  - Lossy data reduction
  - Data is almost random—hard/impossible to compress; <10% lossless
  - N is somewhat *ad hoc* (N=10 for GTS, N=100 for Supernova)
- **ISABELA:**
  - Reduces the data by more than 75%
  - Guarantees user-specified error-bounds
    - Per point error (e.g., **<1%**)
    - Overall correlation (**>0.99***) & NRMSE (**<0.01***)

Metric	FPC	LZMA	ZIP	BZ2	ISABELA	Wavelets	B–splines
Lossless?	Yes	Yes	Yes	Yes	No	No	No
CR_M (%)	-0.55	2.72	1.13	1.11	81.44*	22.51*	0*
Compression (sec.)	0.58	7.01	1.03	3.96	0.93	0.62	0.78
Decompression (sec.)	0.56	1.38	0.49	1.18	1.05	0.58	0.82

EFFECT OF PRECONDITIONER ON UNDERDETRMINED CLASSIFICATION PROBLEM (BENCH)



- **Accuracy increase by 13%-16%**
- **Across different classifiers**
- **On data with <100 samples >*d*=4,000-7,000 dimensions—underdetermined problems**

- **Training time reduced from 1453 sec. (single BBN) to 25 sec. (BENCH)**

- **When applied to seasonal hurricane prediction (*d*>35K), correlation with observed improved: 0.64 → 0.92-0.96**

Classifier	Single classifier	BENCH ensemble
BBN	82.856	97.142
Decision Tree	82.856	95.714
SVM	91.426	97.142

SUMMARY

- **Exascale Analytics = Data-Driven Understanding of Grand Challenge Science Application Problems**
- **Co-design solutions with application scientists**
- **End-to-end data analytics at scale requires different technologies spanning all layers of SW stack**
- **"Optimized" analytical kernels must be customized for different performance metrics and constraints**
- **Preconditioners for data analytics are in very promising but in their infancy**