Report on BDEC (Big Data and Extreme Computing) 2013 and Associated Recent Activities

Satoshi Matsuoka
Tokyo Inst. of Technology

ビッグデータと”Extreme Computing”報告のサマリ(1)

- 以下最近のDoEやNSFの報告会合、更に松岡の研究より抜粋
- 科学技術は種々の分野で観測・シミュレーションのデータ及びその処理要求が爆発、さらに処理形態がインタラクティブ化しており、<u>ビッグデータ技術のメインの適用分野</u>と言える。
- しかし、現状のスパコンは計算・バッチ処理中心で、必ずしもそのような流れに適していない。一方IDC・クラウド中心のビジネス系のビッグデータインフラは、性能や適用要件が狭い等の理由で、やはり適していない。
- 将来の科学技術、さらにそのビジネス適用において、ビッグデータ技術とスパコン技術は競合すべきではない；むしろ将来に向けてシステムに対する要求用件は類似性が高く、寧ろ分野融合してスパコン技術がリーダーシップをとることをはかるべきである。

ビッグデータと"Extreme Computing"報告のサマリ(2)

- ビッグデータ技術の科学技術への適用理由は幾つかの話がある
  - 1. Data Intensive Computing
    - Extremely high bandwidth (Bytes/s) requirements for data processing, many HPC apps are data intensive
    - Not necessarily I/O intensive but often so
  - 2. Data Driven (Data Discovery) Computing
    - Bottom up discovery of unknown relationships
    - C.f. Top-down hypothesis validation
  - 3. Interactive Data Computing
    - Real-time query of data for discovery
    - Often associated with data-driven computing

ビッグデータと"Extreme Computing"報告のサマリ(3)

- **IDC型クラウド、Amazon 10万ノード**
  - **コモディティサーバ(CPUはXeonで高速), VMも**
  - **HDD on node, GFSなど、性能より信頼性**
  - **1Gbps/nodeをラック単位(40ノード程度)で10GBpsに集約, 所謂North-South通信が主**
  - **I/Oやネットワークバンド幅が問題外に足りない**
    - **10万ノードの合算ネットワーク性能は高々10Tbps、バイセクションは更に遥かに低い。(数分の一から数百分の1)**
    - **レーテンシもサブミリセカンドから100ms[Panda]**
  - **Hadoop等それらに特化したbig data のSW abstraction**
    - **グラフなど非定型も無理支離[GIM-V]**
- **スパコン、京クラス10万ノード**
  - **CPUはIDCクラウドと近似**
  - **数万規模HDD+大量のStorage Service Server+Lustre/GPFS, 高性能(1Terabyte/s)**
  - **スパコン用専用超高速低レーテンシネットワーク、数百Tbpsバイセクション、レーテンシ2マイクロ秒**
  - **Graph500の上位はスパコンが独占**
  - **しかし、それでもI/Oバンド幅は足りず、また並列ファイルシステムによるI/Oレートの低下等他にも問題満載**
  - **スケジュリングのリアルタイム性にも欠ける**
  - **Hadoopのスパコン上の実装等はあるが、全然ハードが違うIDCを前提としているので全く性能を生かし切れていない**

<u>どちらもNG => 従来型ビジネス系のビッグデータとスパコンの「コンバージェンス」が重要</u>

ビッグデータと”Extreme Computing”報告のサマリ(4)

- Intel VP of Technical Computing Rajeeb Harza Presentation at CUG, May/8/2013, Napa (Extract)
- ビッグデータの本質はモデル化が難しい未知のwhat-ifのデータを知識に転換する探求。脳からSNSから経済から多くのサイエンスまで今後最も技術的にもマーケットとしても重要。従来のエンタプライズ系のマーケットがクラウドに駆逐されたのは他山の石だ。
- ビッグデータ向けのアーキテクチャは今とは思想を変えなくてはならない。しかしながら、現状のHPCと共通の基盤技術は多く、それらを態々ビッグデータ向けに新たに造るのも馬鹿らしく、むしろ統合すべきだ。Intelが超高速ネットワークをCPU統合するのもそれであり、両方の分野にコミットする。
- 今のビッグデータは、実は企業毎にサイロ化されている。なので、ビッグと言っても量はそれほどでもない。ところが、サイロが打破されるとデータ相関が重要になるけど、これは基本N^2のデータ移動だから、高速化には高バンド幅のスパコン処理となり、オンチップネットワークが重要に
  - 松岡注：サイエンスでは昨今のオープンデータ化によって多くの分野でこれが既に起こっている。
- ビッグデータ系は計算量やデータ移動量に応じて、メモリ多階層にユーザレベルでデータを適切に配置できるかが鍵

10^{18}

BDEC

Attendees:
US: 25
Europe: 11
Japan 9

Exec Committee
Pete Beckman
Jean-Yves Berthou
Jack Dongarra
Yutaka Ishikawa
Satoshi Matsuoka
Philippe Ricoux

Charleston, South Carolina, USA, April 30- May 1

BIG DATA AND EXTREME-SCALE COMPUTING

http://www.exascale.org/bdec/

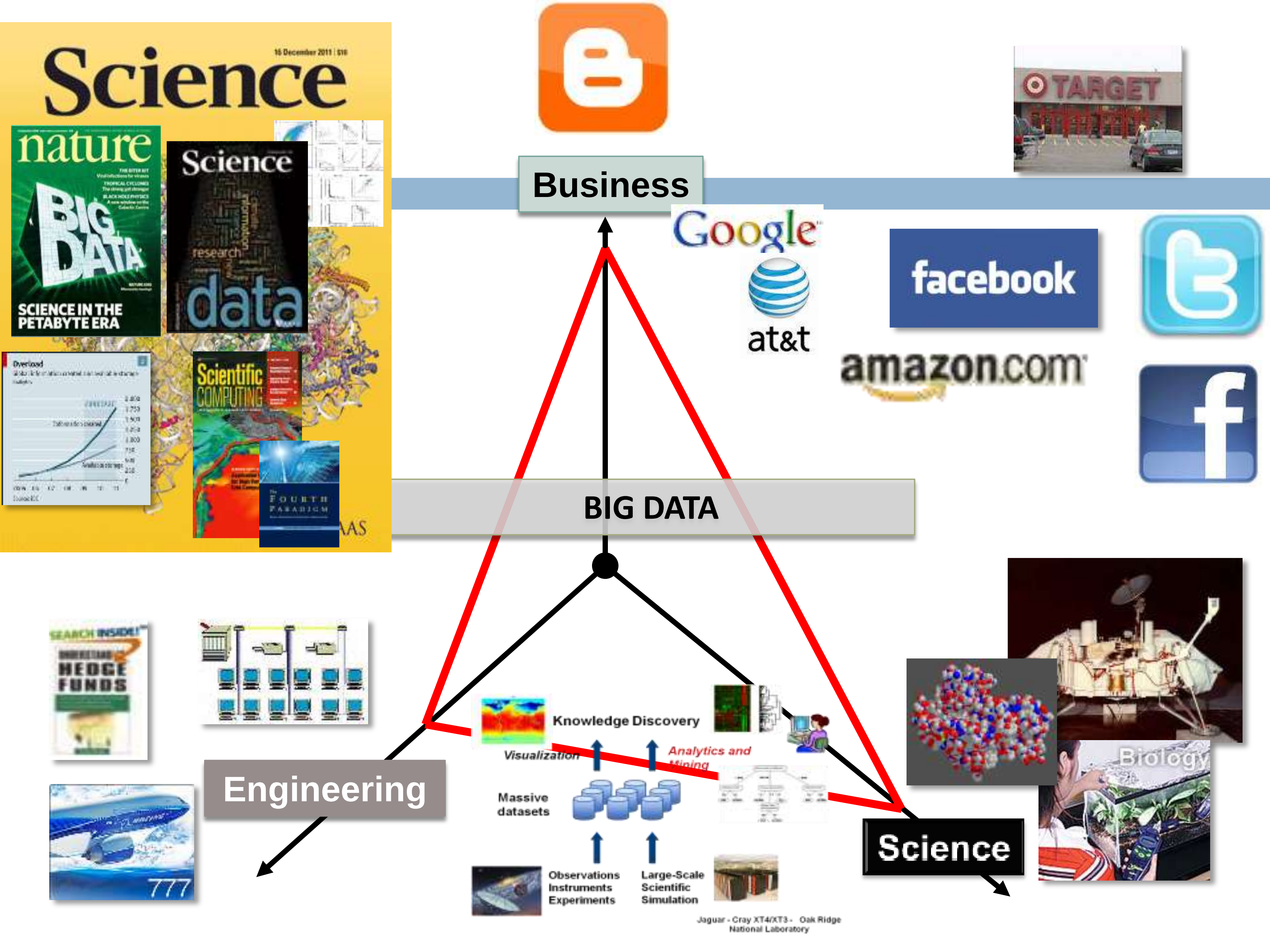
Other Activities April-May 2013 on Big Data

- DoE Anjual “Salishan” Meeting:
  - Apr 22-25, 2013, Salishan, Oregon, USA
  - “Big Data” was THE theme, many DoE and vendor talks
- Cray Users Group Meeting
  - May 6-8, Napa, California, USA
  - Big Data emphasis by Cray and users in many presentations, e.g. Intel Keynote
- IEEE CCGrid2013 (Delft, Netherlands, May 13-16), IEEE IPDPS2013 (Boston, MA, May 20-24)
  - Emphasis on Big Data, including keynotes Dan Reed (NCSA=>MS=>Iowa U)@CCGrid, VMWare@IPDPS, many papers on Hadoop, graphs, etc.

Big Data's Biggest Needs– Deep Analytics for Actionable Insights

Alok Choudhary
John G. Searle Professor
Dept. of Electrical Engineering and Computer Science
and Professor, Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu

National Science Foundation
WHERE DISCOVERIES BEGIN
ACKNOWLEDGEMENTS
U.S. DEPARTMENT OF ENERGY



Business
BIG DATA
Engineering
Science
Science
nature
BIG DATA
SCIENCE IN THE PETABYTE ERA
Science
data
Scientific COMPUTING
FOURTH PARADIGM
TARGET
Google
at&t
facebook
amazon.com
HEDGE FUNDS
777
Knowledge Discovery
Visualization
Analytics and Mining
Massive datasets
Observations
Instruments
Experiments
Large-Scale
Scientific
Simulation
Jaguar - Cray XT4/XT3 - Oak Ridge National Laboratory
Biology

“Data intensive” vs “Data Driven”

Data Intensive (DI)

- Depends on the perspective
  - Processor, memory, application, storage?
- An application can be data intensive without (necessarily) being I/O intensive

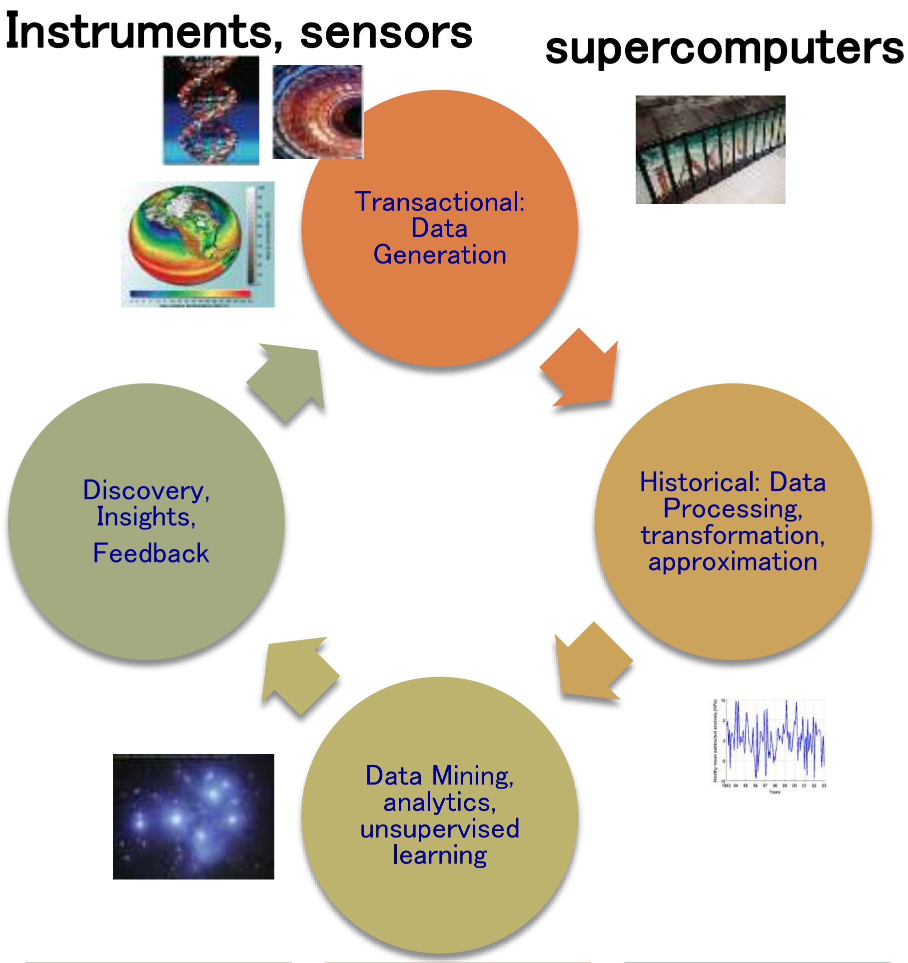
Data Driven (DD)

- Operations are driven and defined by data
  - BIG analytics
    - Top-down query (well-defined operations)
    - Bottom up discovery (unpredictable time-to-result)
  - BIG data processing
  - Predictive modeling
- Usage model further differentiates these
  - Single App, users
  - Large number, sharing, historical/temporal

Very few large-scale applications of practical importance are NOT Data

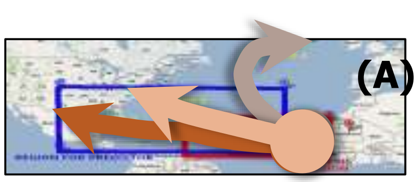
In Extreme Scale Science domain, we typically focus on “Transactional” thinking

Knowledge Discovery Life-Cycle: Transactional to Relationships – Current to Historical

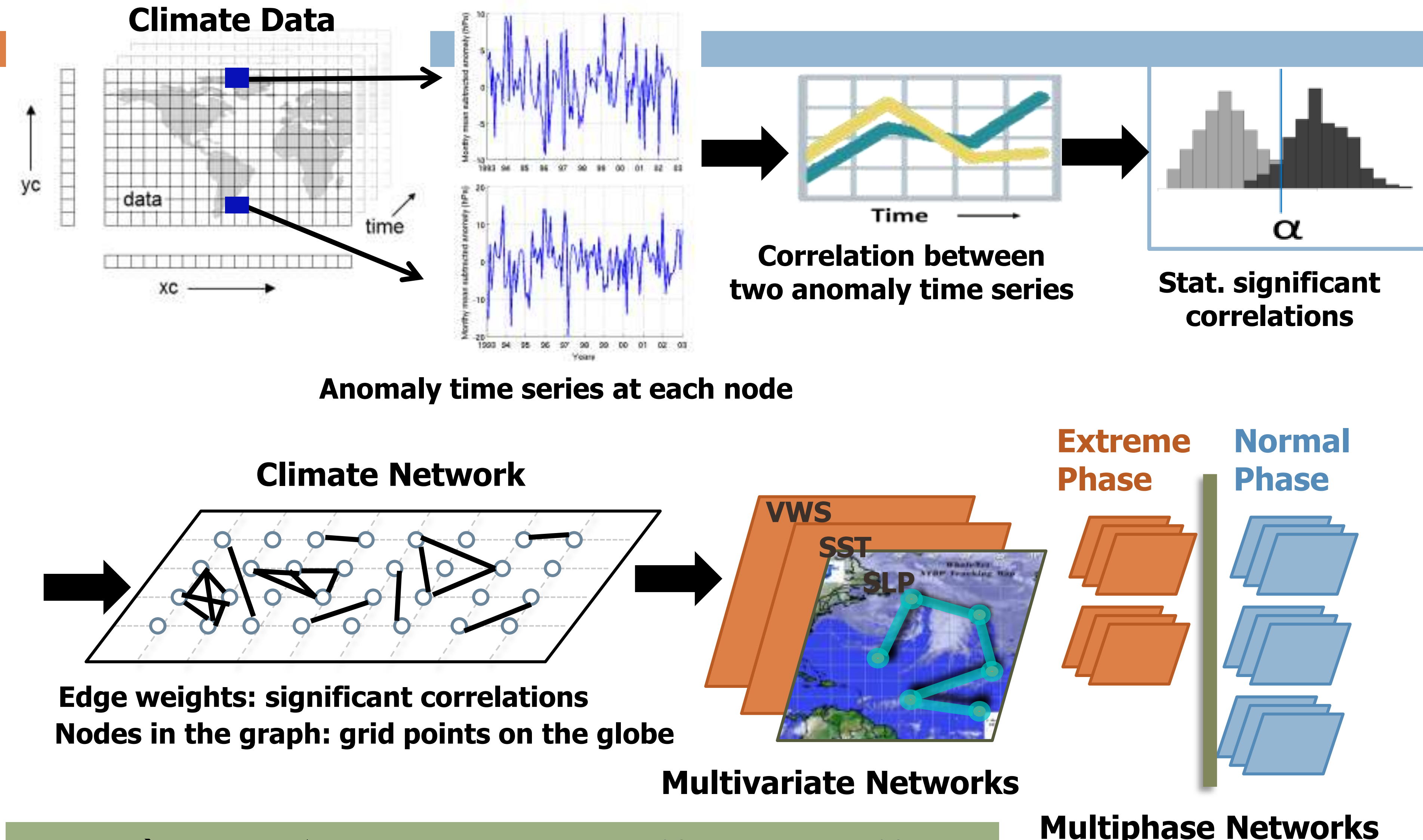


- Data Management
- Data Reduction, Query
- Data Visualization
- Data Sharing

Historical data → Learning Models → Trigger/questions → Predict

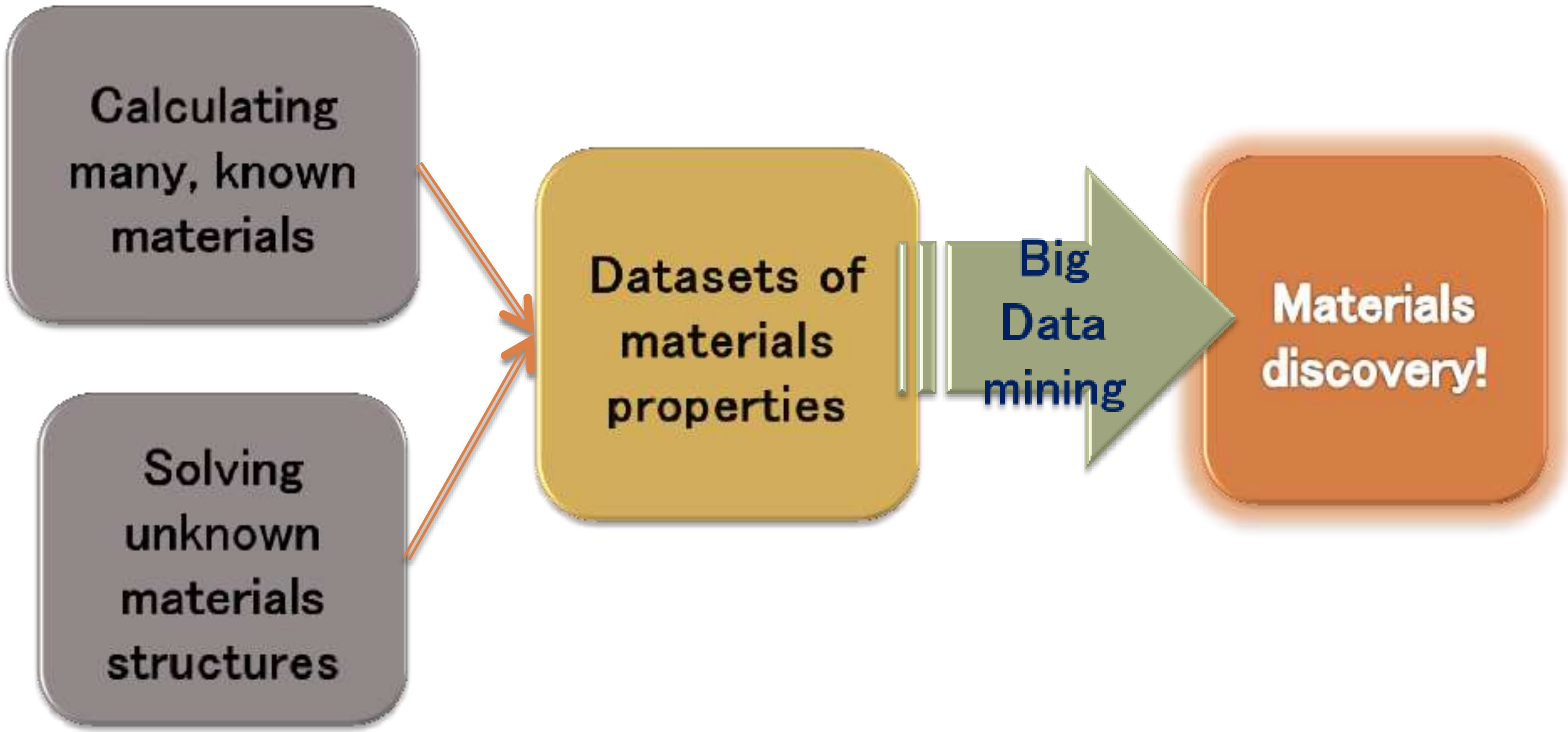


From multi-dimensional data analytics to relationship mining

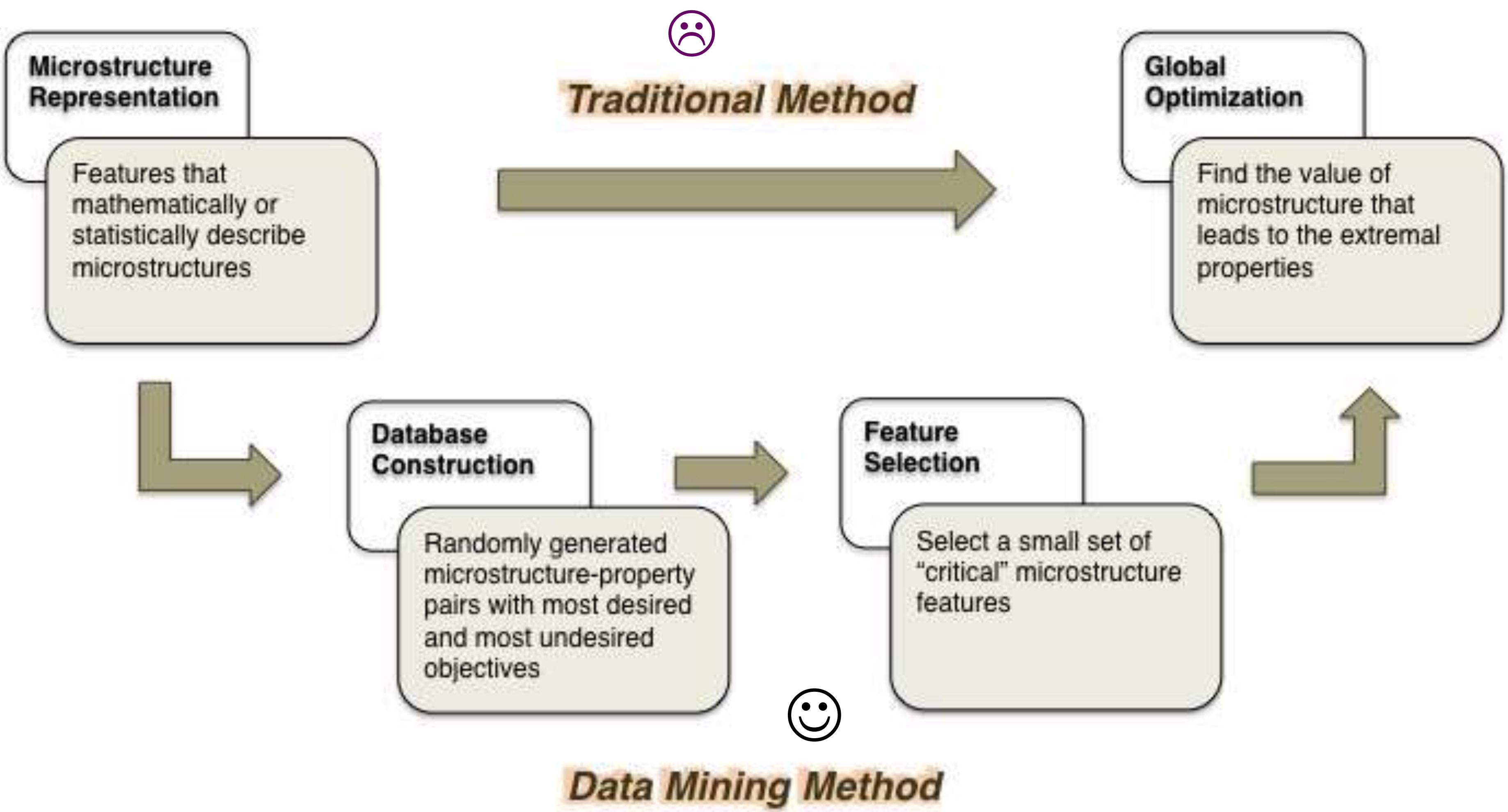


CMIP3 → CMIP5 => Climate BIG DATA : 10s of TBs to 10s of PBs

Discovery of stable compounds



Structure–Property Optimization – Try optimization for 10^3 dimensions

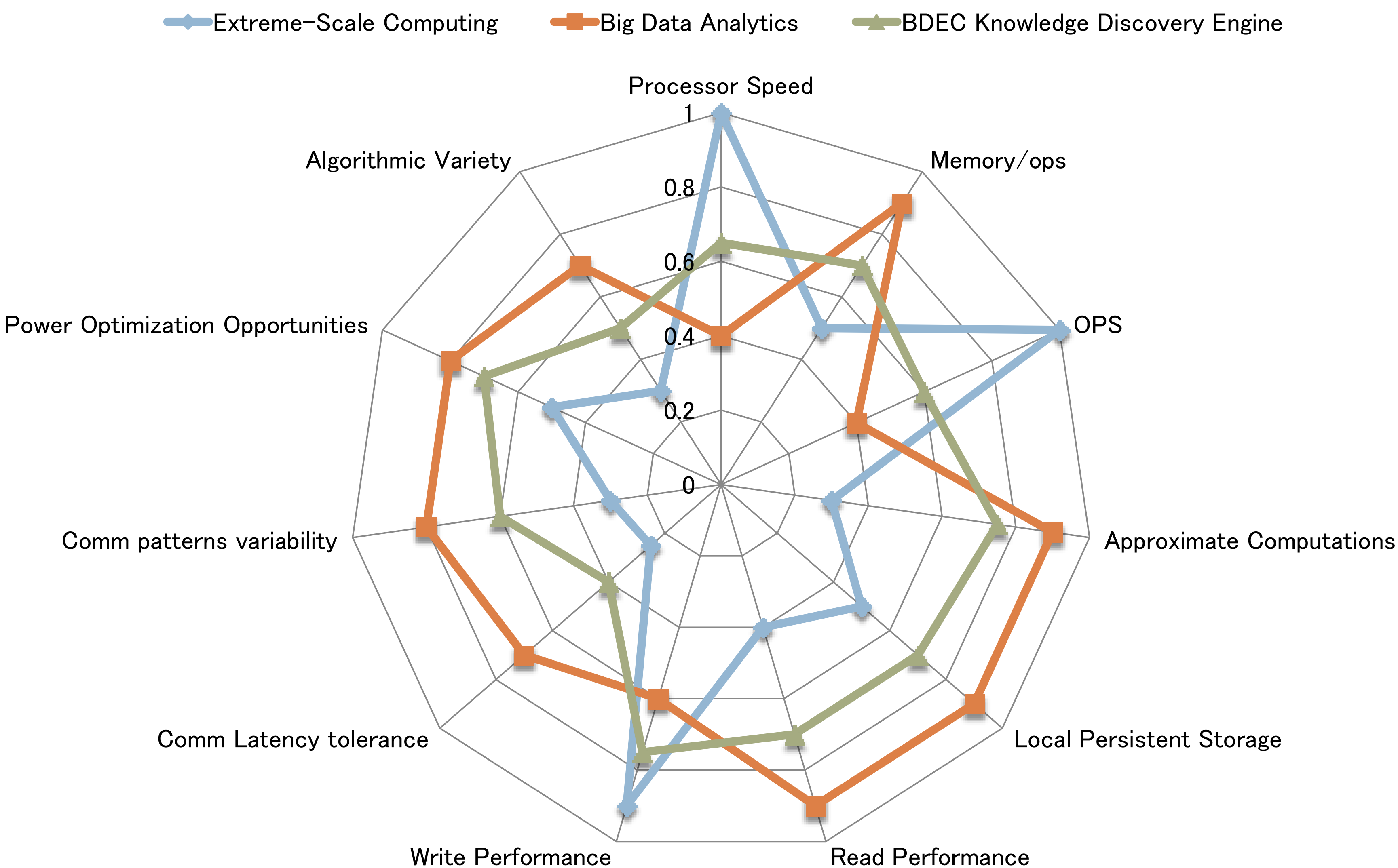


Right Computing infrastructure?
What characteristics do typical analytics functions have?

Parameter[†]	Benchmark of Applications				
	SPECINT	SPECFP	MediaBench	TPC-H	MineBench
Data References	0.81	0.55	0.56	0.48	1.10
Bus Accesses	0.030	0.034	0.002	0.010	0.037
Instruction Decodes	1.17	1.02	1.28	1.08	0.78
Resource Related Stalls	0.66	1.04	0.14	0.69	0.43
CPI	1.43	1.66	1.16	1.36	1.54
ALU Instructions	0.25	0.29	0.27	0.30	0.31
L1 Misses	0.023	0.008	0.010	0.029	0.016
L2 Misses	0.003	0.003	0.0004	0.002	0.006
Branches	0.13	0.03	0.16	0.11	0.14
Branch Mispredictions	0.009	0.0008	0.016	0.0006	0.006

[†] The numbers shown here for the parameters are values per instruction

Extreme Computing + Big Data Analytics = BDEC Knowledge Discovery Engine



Big Data, Big Compute, Big Interaction Machines for Future Biology

Rick Stevens

stevens@anl.gov

Argonne National Laboratory

The University of Chicago

BD Usage Models Differ from EC

Big Data

- Continuous access require based on data generation/submission rates
- CPU time, I/O and data volume all important
- Data products typically used in future computations via an integration or pipeline
- Data products made available for external users and curated over time

Extreme Compute

- Batch oriented access based on allocations for specific projects
- Mostly CPU time centric
- Output not necessarily used in future runs but often significant time used for visualization
- Output generally (but not always) used “privately” and rarely curated

Policies Need to be Different

- Long term (many years) access commitment at a continuous or increasing level of service
- Support for persistent services
- Storage allocation that grows over time
- Rich software environment with high-performance database support
- Mechanism to publish the data to a community
- Archival support for data, links and citations

Convergence

- Ideal Environment
  - Interactive parallel prototyping environment
  - Seamless scale up to production (10^3x-10^6x)
  - Integrated platform for analysis and simulation
  - Same platform for publishing
  - Persistent data regions in memory
  - Programming language support for data analysis
  - Large-scale interactive computing
  - Seamless visualization and sharing


MIRA
DataDirect Networks

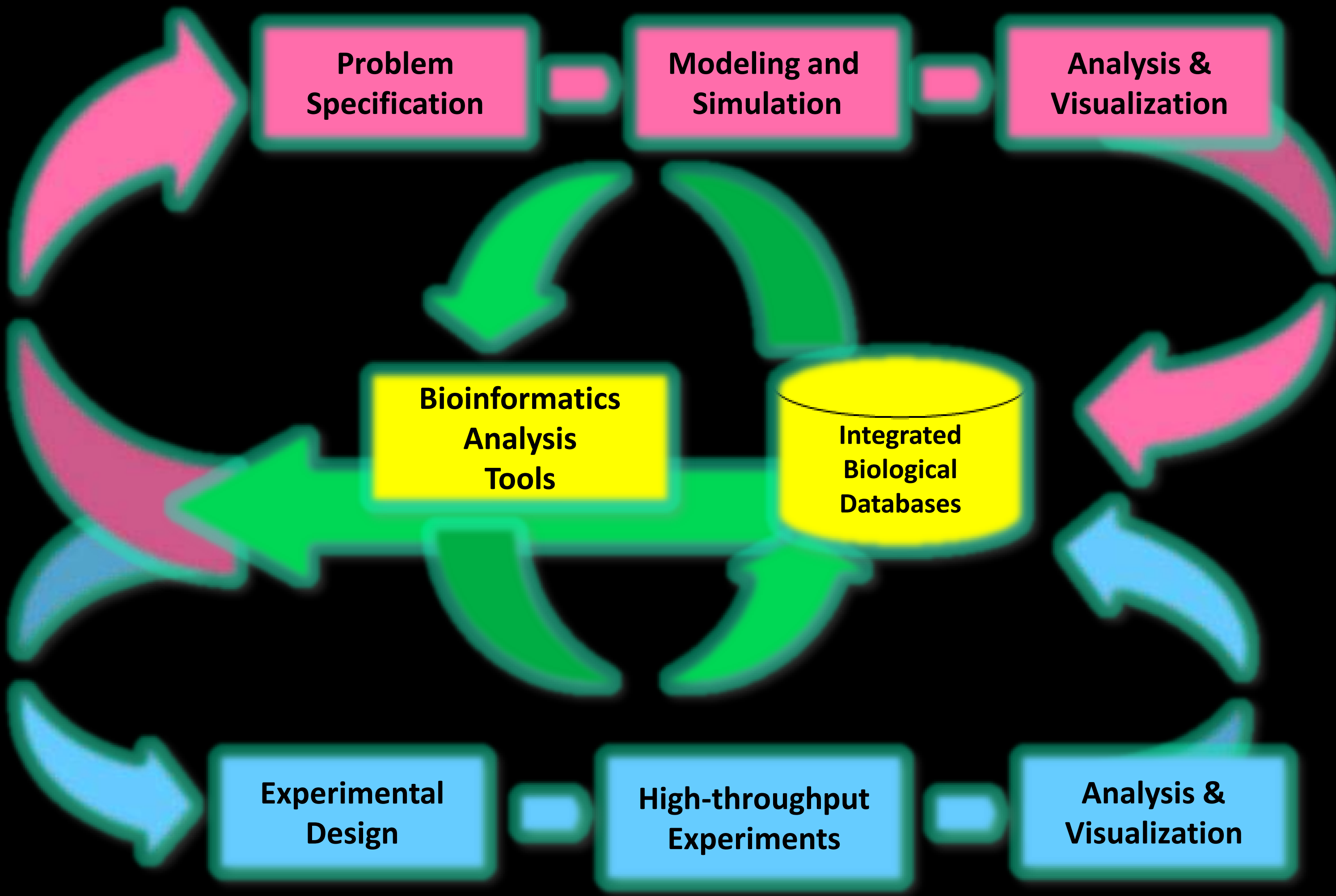
Magellan: Our OpenStack Private Cloud for Systems Biology

What do we want to do with Data?

- Generate
- Process
- Analyze
- Annotate
- Visualize
- Understand
- Share
- Publish
- Curate
- Archive
- Integrate
- Move
- Search
- Preserve
- Model
- Compare

GREEN is Interactive

Converging View of Modeling, Simulation, Experiment, Data and Bioinformatics



Big Data Challenges for Bioinformatics

- New types of methods and new algorithms
  - From O(N3) $\Rightarrow$ $O(N^2)$ $\Rightarrow$ $O(N \log N)$ $\Rightarrow$ $O(N)$ $\Rightarrow$ $O(K)$
  - Non-alignment methods and streaming
- New types of Infrastructure bringing biological data and computing together
  - Users need to have an environment where they don't need to move the data to work
- Ability to share methods, protocols, tools and insights leveraging social networks
  - Enable the best methods to win regardless of where they come from

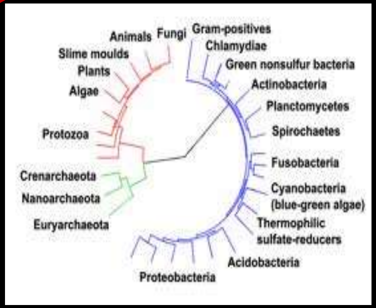
Sequencing the Environment

Metagenomic data collection

Collecting samples

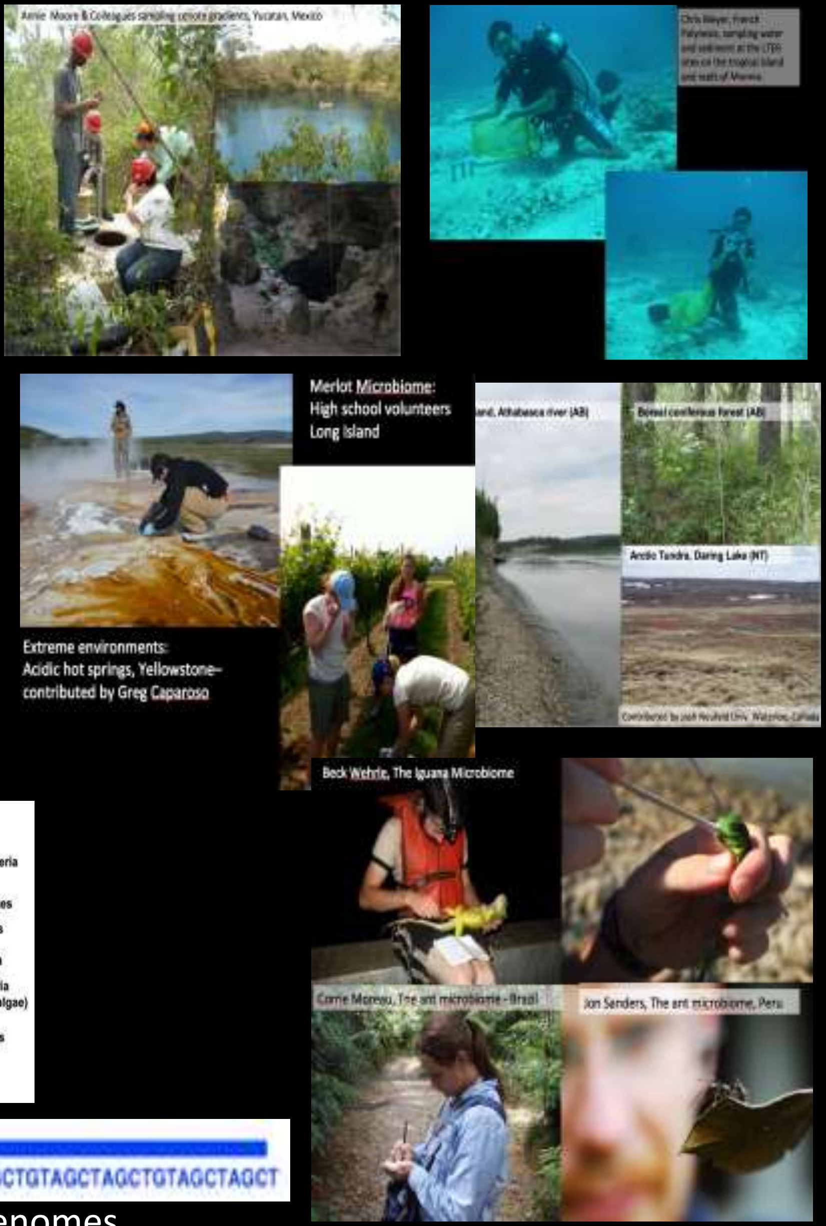
Sequencing

Sequence fragments



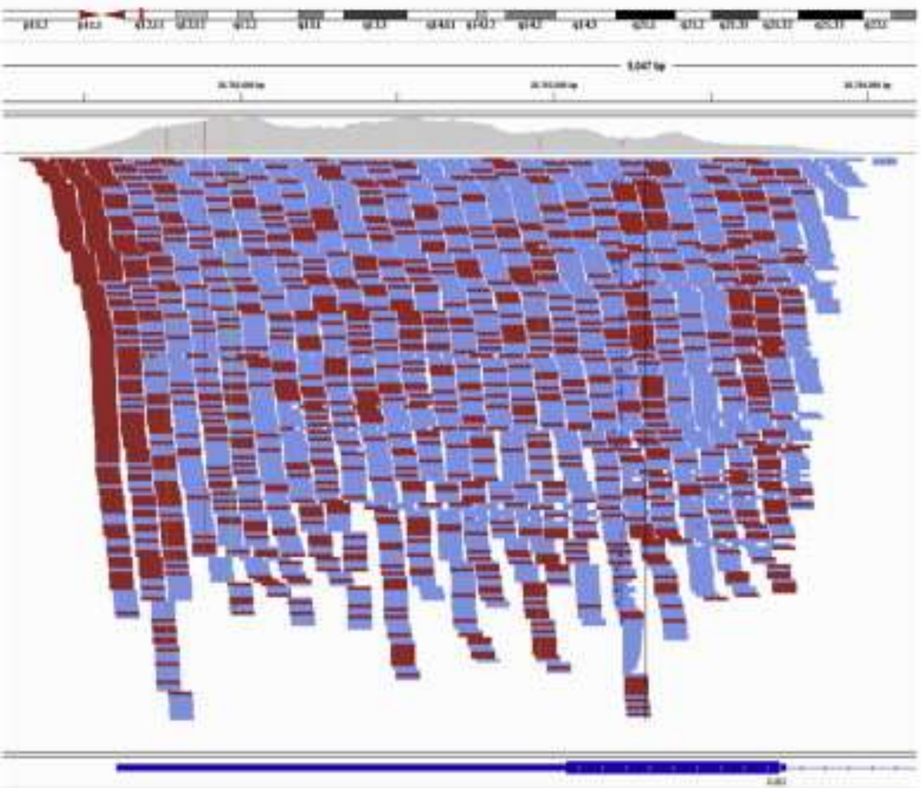
Associating fragments to taxonomical groups

Assembly of most abundant microbes into complete genomes

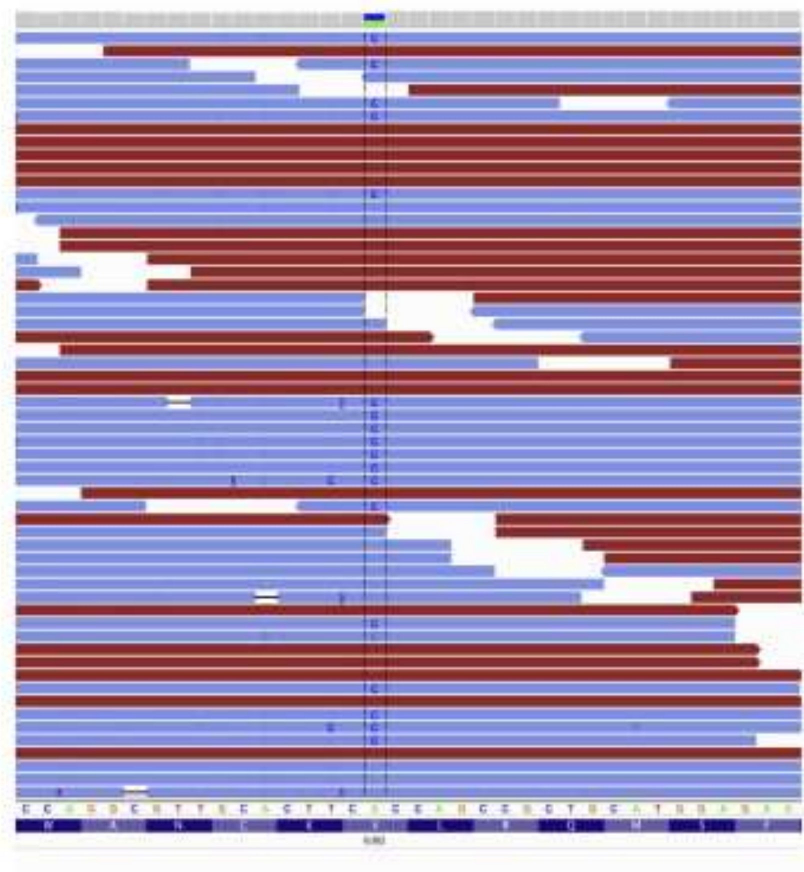


a Illumina HiSeq
Generate Millions of Sequencing Reads

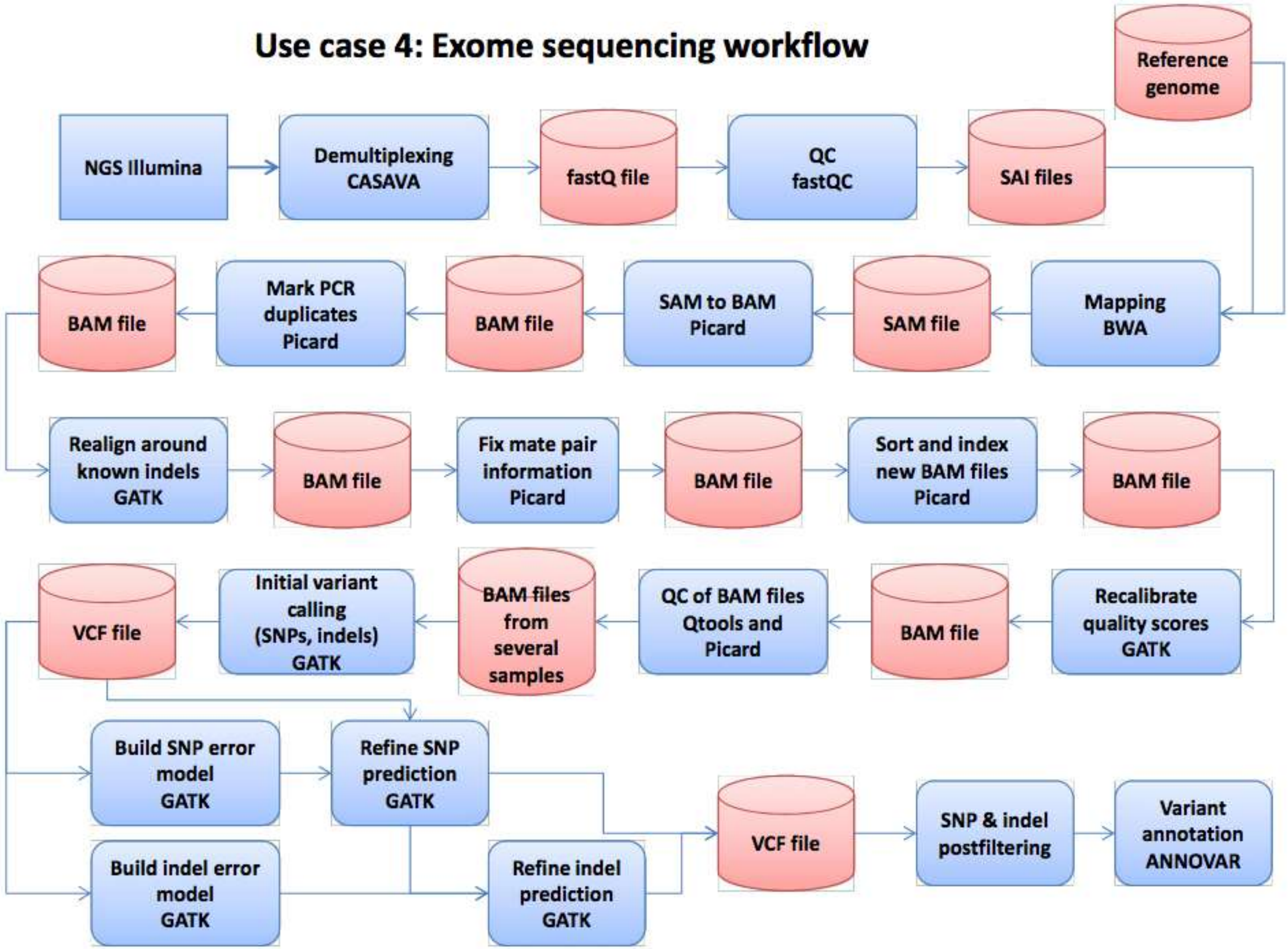
b Raw Sequencing Read Output



c Align Millions of Sequencing Reads to the Reference Genome



d Identify and Annotate Variant:
GJB2 Val167Met carrier


Use case 4: Exome sequencing workflow
Reference genome
NGS Illumina
Demultiplexing CASAVA
fastQ file
QC fastQC
SAI files
BAM file
Mark PCR duplicates Picard
BAM file
SAM to BAM Picard
SAM file
Mapping BWA
Realign around known indels GATK
BAM file
Fix mate pair information Picard
BAM file
Sort and index new BAM files Picard
BAM file
VCF file
Initial variant calling (SNPs, indels) GATK
BAM files from several samples
QC of BAM files Qtools and Picard
BAM file
Recalibrate quality scores GATK
Build SNP error model GATK
Refine SNP prediction GATK
VCF file
SNP & indel postfiltering
Variant annotation ANNOVAR
Build indel error model GATK
Refine indel prediction GATK

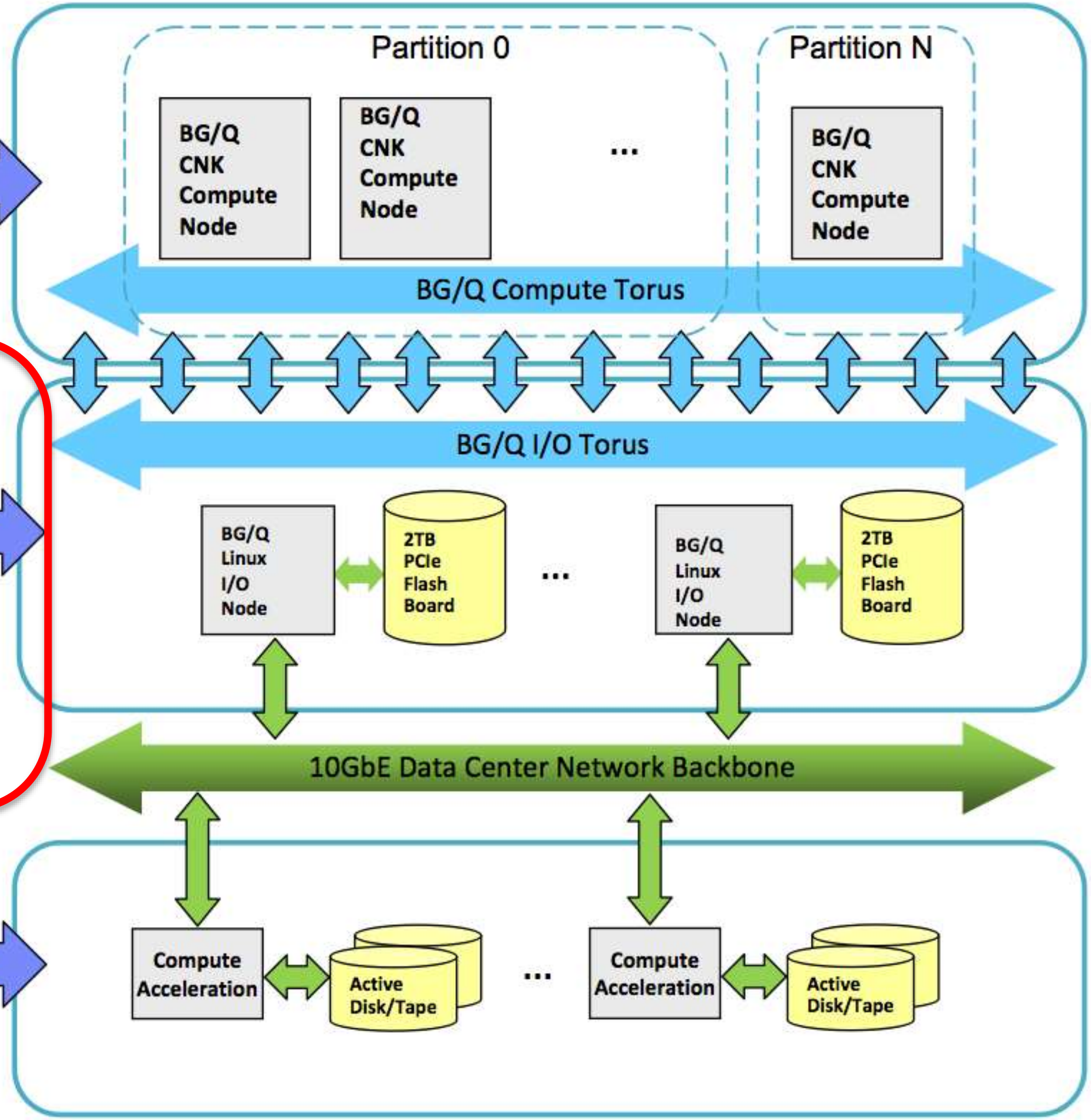
Data-centric Computing Using BG/Q Active Storage

IBM

High Compute Density
- BG/Q compute fabric 1k – 100k nodes
- DRAM memory
- 5D Compute Node Torus
- CNK, ZeptoOS, FuseOS
- I/O Links to D1 Layer (4:1 ratio)

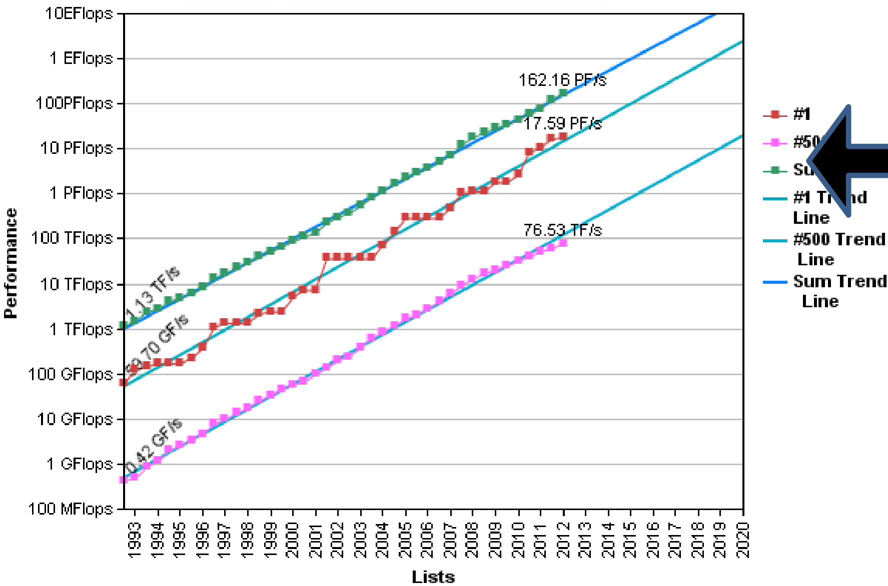
Active Storage
- 8 – 4096 Linux BG/Q I/O Nodes
- DRAM + 2TB SLC Flash per Node
- 2 GBps bandwidth to storage
- GPFS / KV services
- I/O links to each node (4GBps/node)
- All to All Comm. Via I/O Torus
- DB2, Infosphere Streams, Hadoop, MVAPICH, SLURM

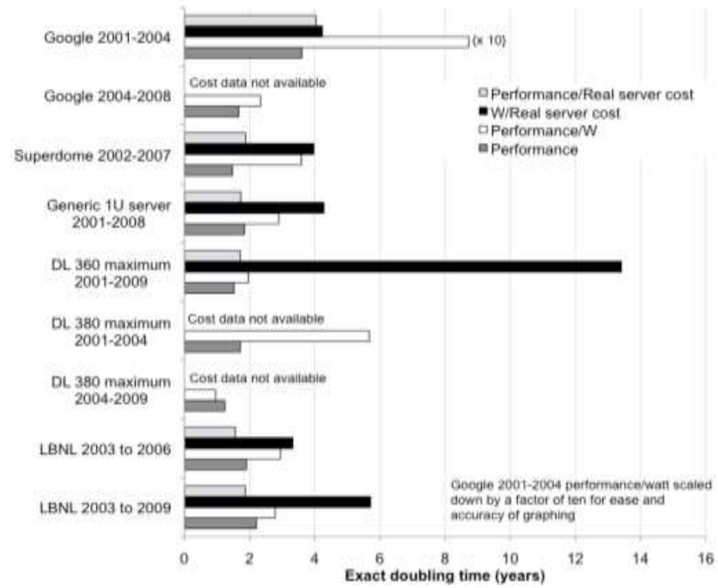
Data Center Storage
- GPFS file system
- External Disk Controller Racks



今後Convergenceをどうするか？

But how do we achieve “convergence” at future extreme scale?





Source: **Assessing trends over time** in **performance, costs**, and **energy use** for **servers**, Intel, 2009.

HPC: x1000 in 10 years

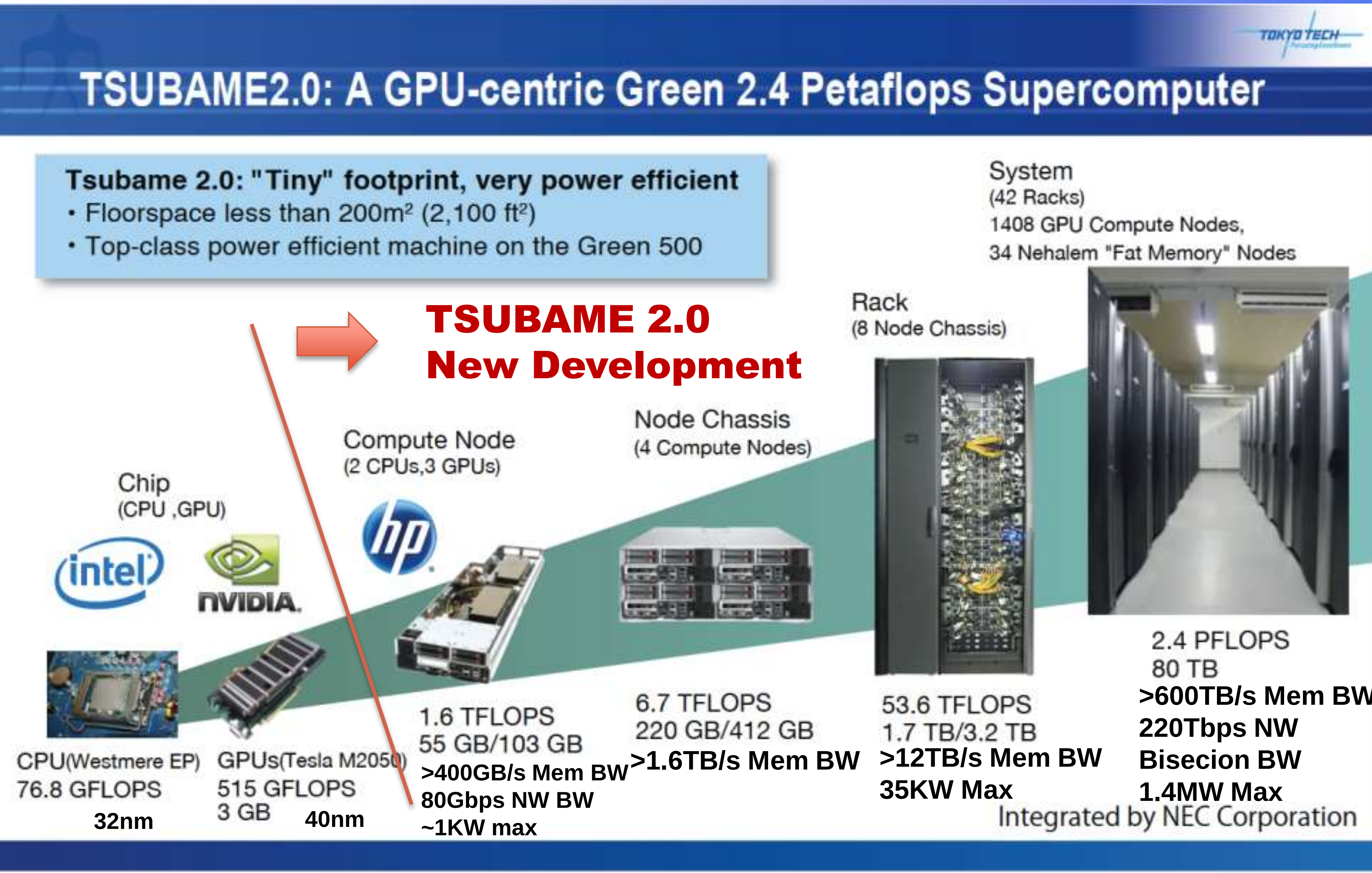
CAGR ~= 100%

IDC: x30 in 10 years

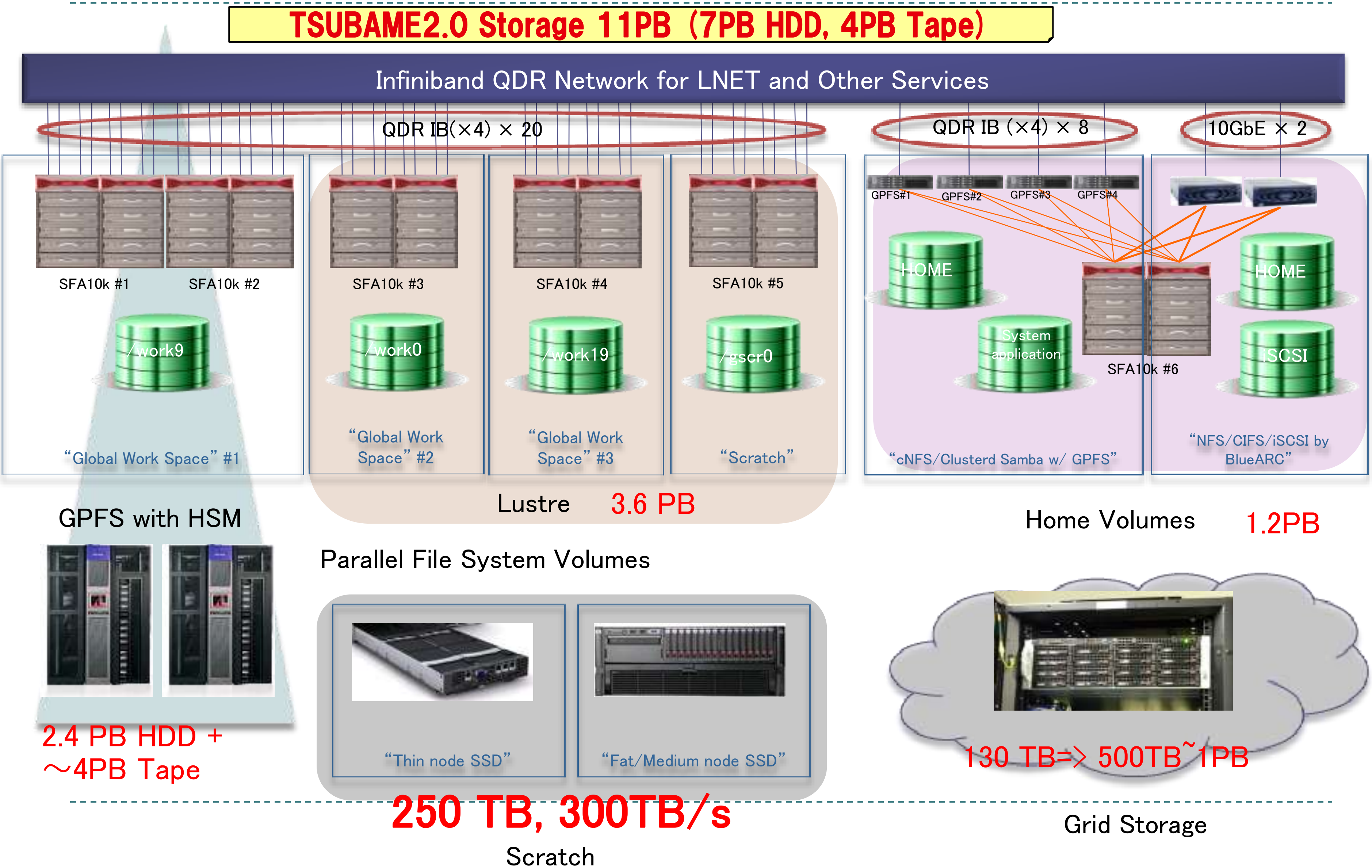
Server unit sales flat (replacement demand)

CAGR ~= 30-40%

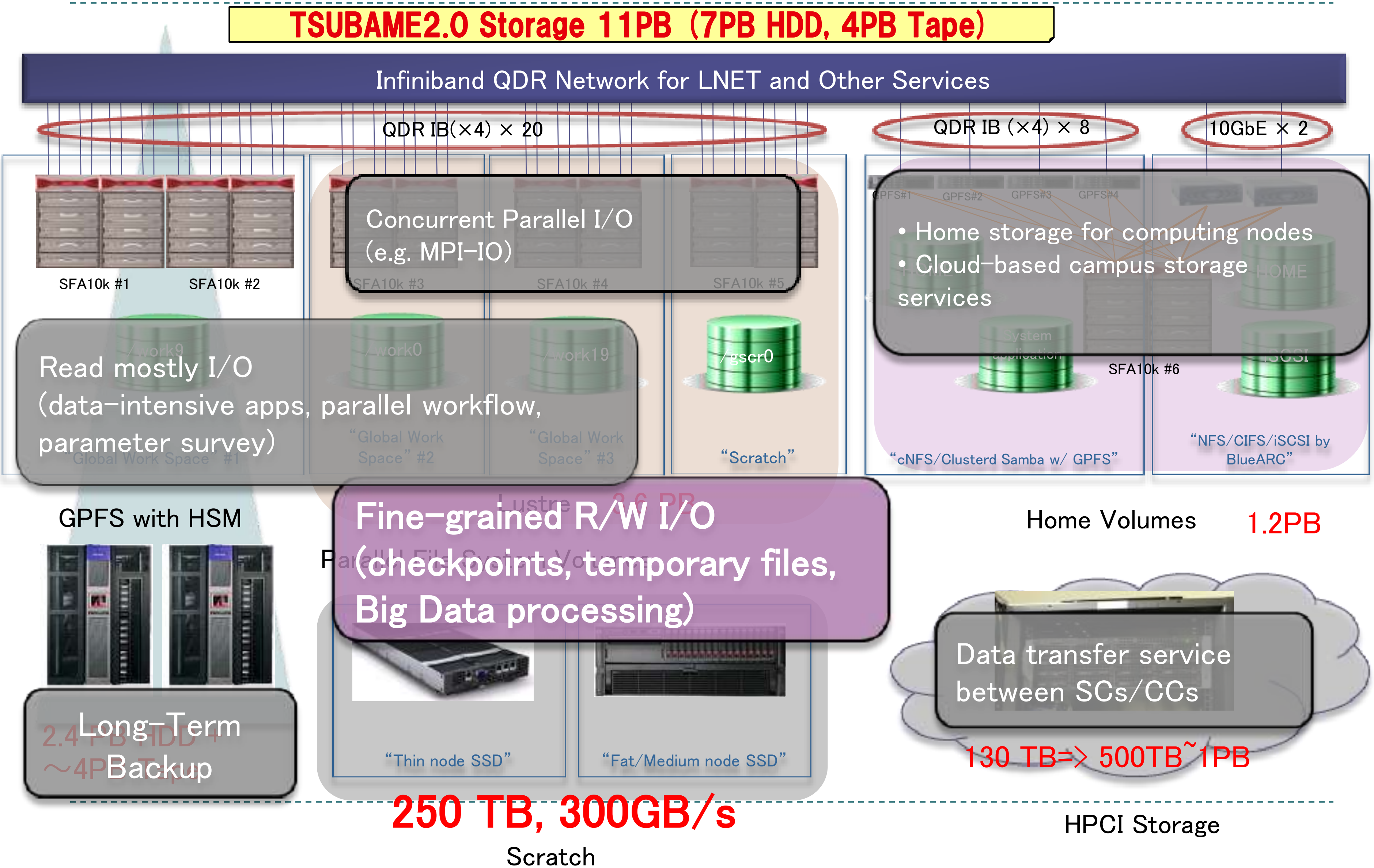
TSUBAME2.0 Nov. 1, 2010
“The Greenest Production Supercomputer in the World”



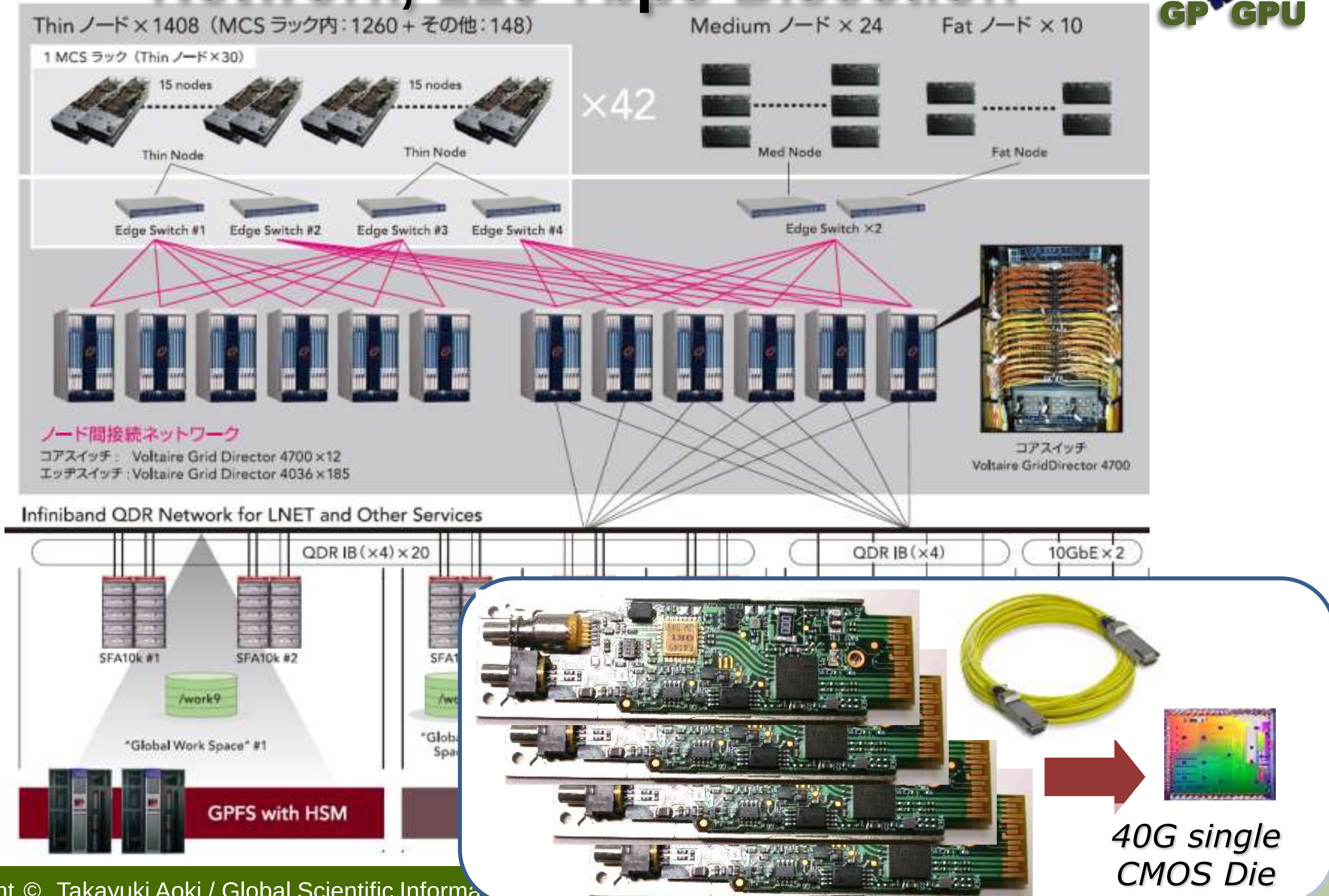
TSUBAME2.0 Storage Overview



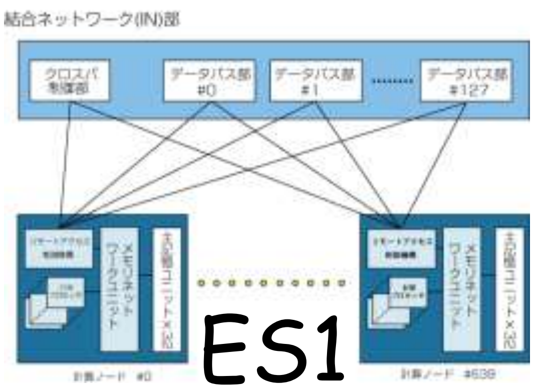
TSUBAME2.0 Storage Overview



Full Bisection Multi-Rail Optical Network, 220 Tbps Bisection

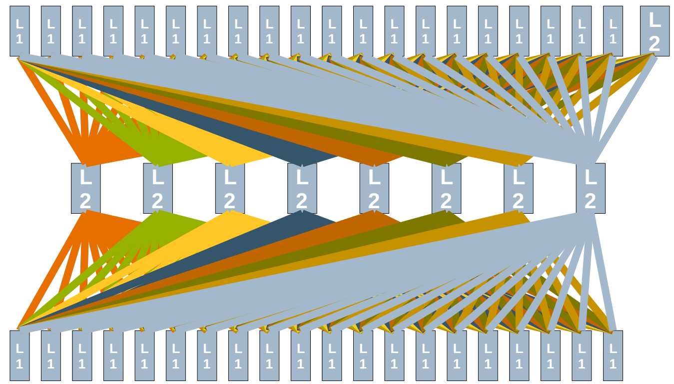


Comparing the Networks



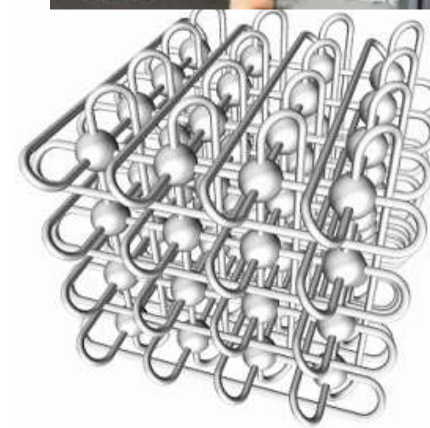
ES1

12.8GB/s Link
5us latency
Full Crossbar
6~7TB/s Bisection BW
3000km Copper



TSUBAME2.0

IB QDRx2 7.5GB/s Node
2us latency
Oversubscribed Full Bisection Fat Tree
~20TB/s Bisection BW
100km DFB/Single Mode Fiber



K Computer

5GB/s Link
5us latency
6-D Torus
~30TB/s (???) Bisection BW
1000km Copper

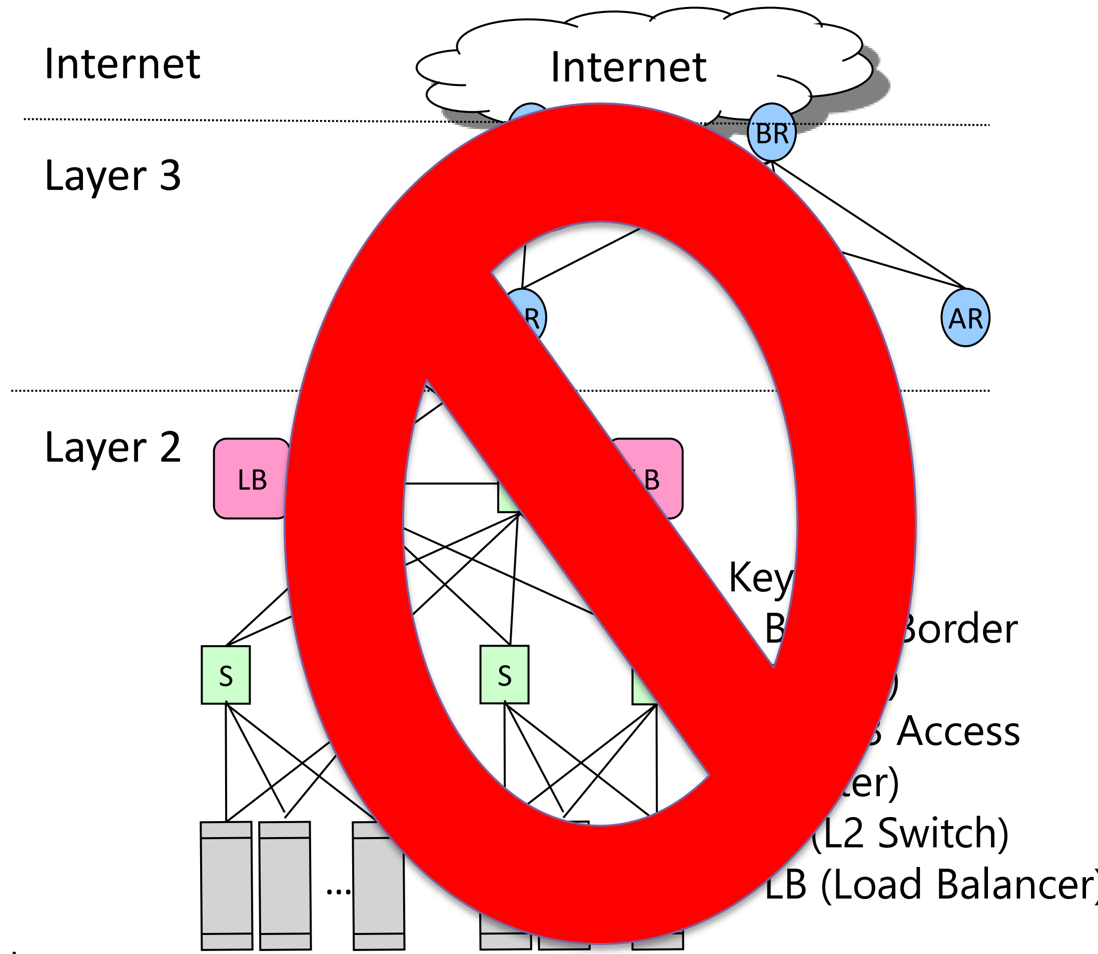
But what does “220Tbps“ mean?

Global IP Traffic, 2011-2016 (Source Cicso)							
	2011	2012	2013	2014	2015	2016	CAGR 2011-2016
By Type (PB per Month / Average Bitrate in Tbps)							
Fixed Internet	23,288	32,990	40,587	50,888	64,349	81,347	28%
	71.9	101.8	125.3	157.1	198.6	251.1	
Managed IP	6,849	9,199	11,846	13,925	16,085	18,131	21%
	21.1	28.4	36.6	43.0	49.6	56.0	
Mobile data	597	1,252	2,379	4,215	6,896	10,804	78%
	1.8	3.9	7.3	13.0	21.3	33.3	
Total IP traffic	30,734	43,441	54,812	69,028	87,331	110,282	29%
	94.9	134.1	169.2	213.0	269.5	340.4	

TSUBAME2.0 Network has TWICE the capacity of the Global Internet, being used by 2.1 Billion users

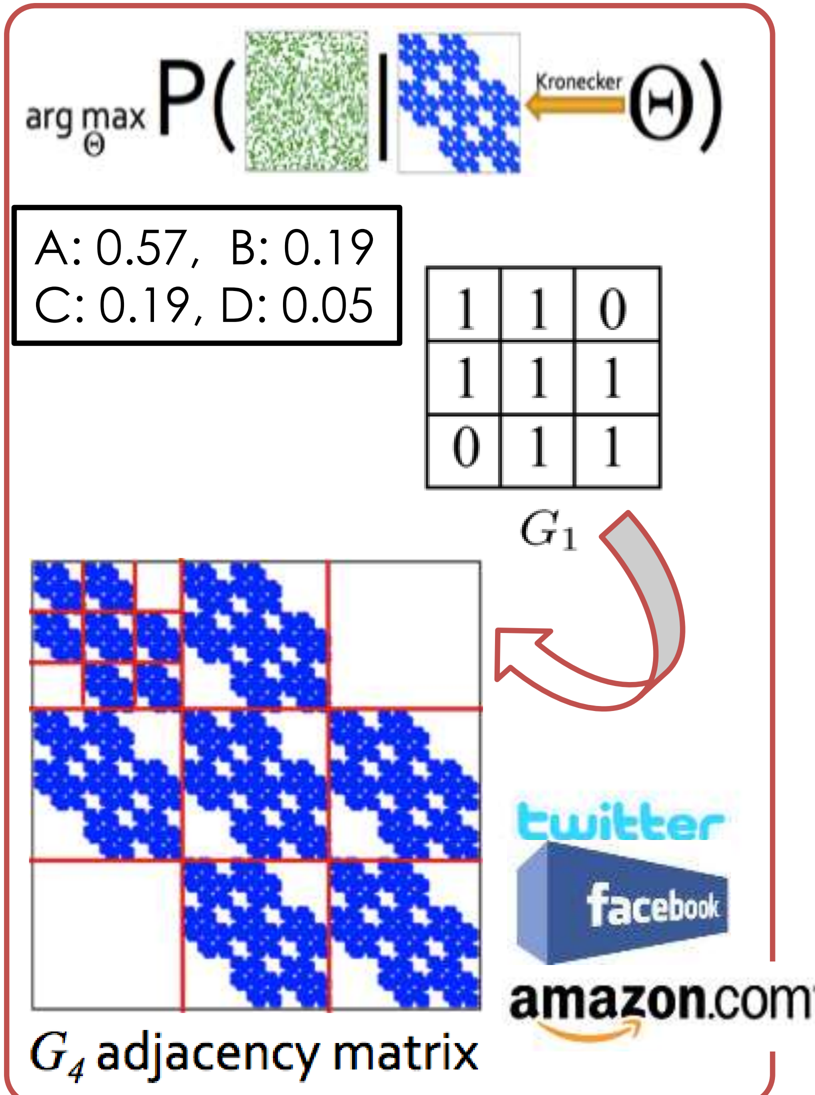
Five years ago, data center networks were here (slide from Dan Reed@MS->Iowa)

- Historical hierarchical data center network structure
  - (Mostly) driven by economics
  - (Partially) driven by workloads
  - Performance limited
- Now moving to “flat” *(sound familiar?)*
  - From N-S to E-W
- Challenges (then and now)
  - Configuration and testing
  - Monitoring and resilience
  - Service demand variance
    - Workload redistribution
    - Service drain times
  - Compatibility (see IPv6 transition)
  - LAN/WAN separation
    - Data islands, geo-resilience and scale
  - Performance and cost, Cost, COST
    - *Did I mention cost?*



Graph500 “Big Data” Benchmark

Kronecker graph



- The benchmark is ranked by so-called **TEPS (Traversed Edges Per Second)** that measures the number of edges to be traversed per second by searching all the reachable vertices from one arbitrary vertex with each team’s optimized BFS (Breadth-First Search) algorithm.

November 15, 2010

Graph 500 Takes Aim at a New Kind of HPC

Richard Murphy (Sandia NL => Micron)

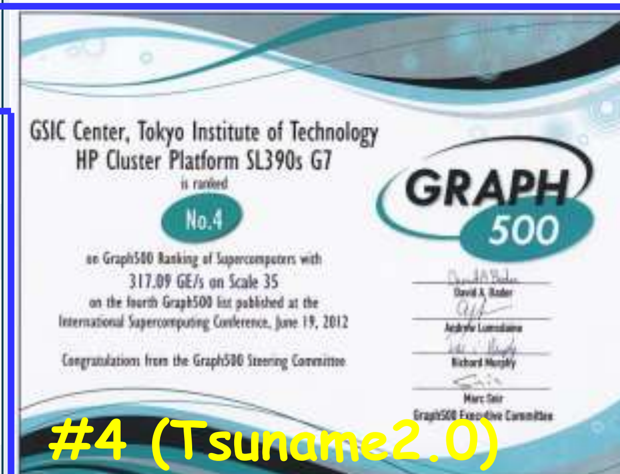
“ the goal of the Graph 500 benchmark is to measure the performance of a computer solving a large-scale "informatics" problem…(for) cybersecurity, medical informatics, data enrichment, social networks, and symbolic networks.”

“ **I expect that this ranking may at times look very different from the TOP500 list. Cloud architectures will almost certainly dominate a major chunk of part of the list**, and we may find that some exotic architectures dominate the top.”

The 4th Graph500 List (Jun2012) TSUBAME #4 w/GPUs

Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology

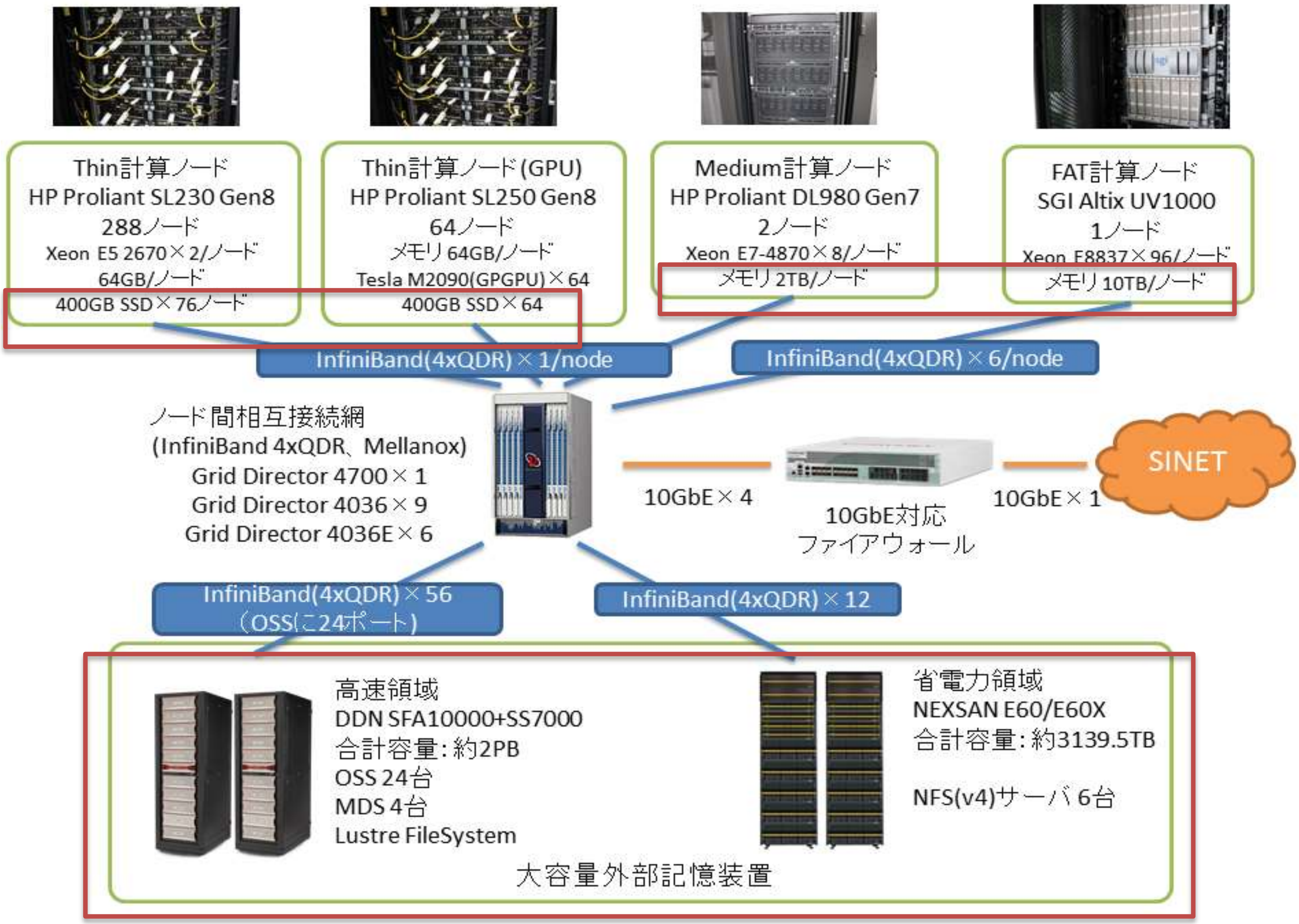
Rank	Installation Site	Machine	Number of nodes	Number of cores	Problem scale	GTEPS
1	DOE/SC/Argonne National Laboratory	Mira/BlueGene/Q	32768	524288	38	3541.00
1	LLNL	Sequoia/Blue Gene/Q	32768	524288	38	3541.00
2	DARPA Trial Subset, IBM Development Engineering	Power 775, POWER7 8C 3.836 GHz	1024	32768	35	508.05
3	Information Technology Center, The University of Tokyo	Oakleaf-FX (Fujitsu PRIMEHPC FX 10)	4800	76800	38	358.10
4	GSIC Center, Tokyo Institute of Technology	TSUBAME	1366	16392	35	317.09
5	Brookhaven National Laboratory	BLUE GENE/Q	1024	16384	34	294.29
6	DOE/SC/Argonne National Laboratory	Vesta/BlueGene/Q	1024	16384	34	292.36
7	NASA-Ames / Parallel Computing Lab, Intel Labs	Pleiades - SGI ICE-X, dual plane hypercube FDR infiniband, E5-2670 "sandybridge"	1024	16384	34	270.33
8	NERSC/LBNL	XE6	4817	115600	35	254.07
9	NNSA and IBM Research, T.J. Watson	NNSA/SC Blue Gene/Q Prototype II	4096	65536	32	236.00



#4 (Tsuname2.0)

Watch out for the new "Green Graph 500" @ISC13

遺伝研DDBJスパコン
– Tsubame2.0の「ビッグデータ」向け仕様 –



Large-Scale Metagenomics
[Akiyama et. al. Tokyo Tech.]

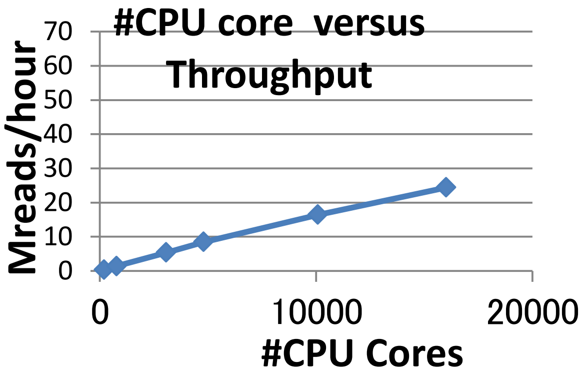
Combined effective use of GPUs and SSDs and 200Tbps Interconnect on TSUBAME2.0.

Metagenome analysis: study of the genomes of uncultured microbes obtained from microbial communities in their natural habitats

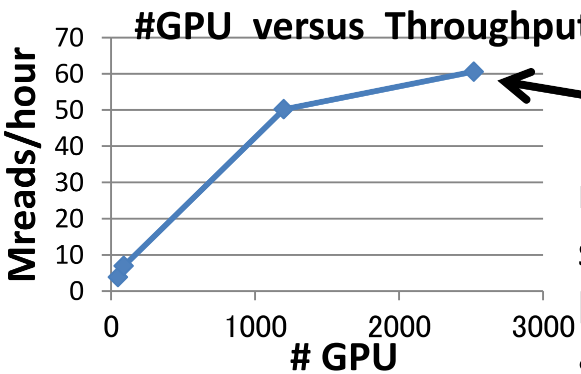
Collecting bacteria in soil

Results on TSUBAME2.0

BLASTX: 24.4M/hour with 16K cores



GHOSTM: 60.6M/hour with **2520 GPUs**



It would be more scalable with larger data sets

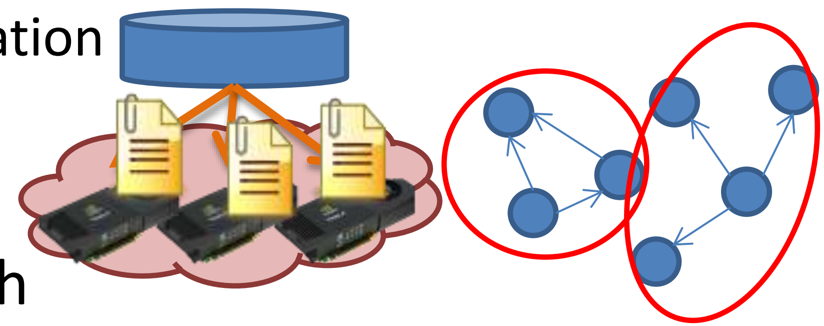
Two homology search tools are available:

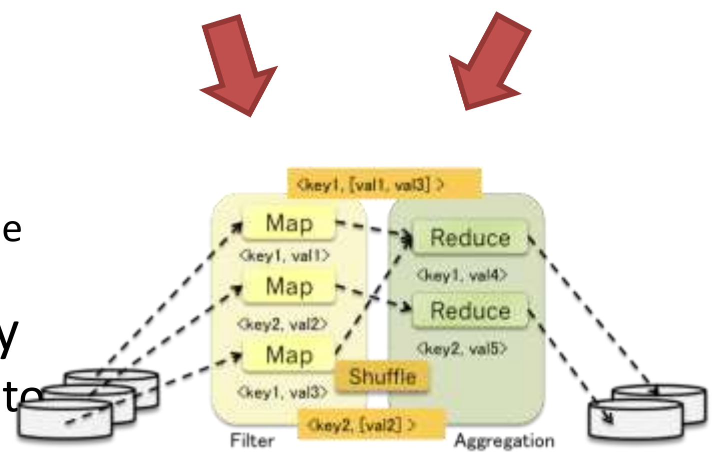
1) BLASTX, standard software on CPUs
2) GHOSTM, our GPU-based fast software compatible with BLASTX

Data: 224million DNA reads(75b) /set
Pre-filtering: reduces to 71M reads
Search: 71M DNA vs. NCBI nr-aa DB (4.2GB)

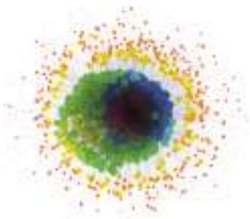
Multi GPU Implementation with Reduction of Data Transfer using Graph Cut [IEEE CCGrid13]

- Investigation of effect of GPU to MapReduce type graph algorithm
  - Comparison with existing implementation
    - Existing CPU implementation
    - Optimized implementation not using MapReduce
- Handling extremely large-scale graph
  - Increase amount of memory using Multi GPU
    - Reduce amount of data transfer
      - As one of the solution, Partition the graph as preprocessing and reduce amount of inter-node data transfer on Shuffle
  - Utilize local storage in addition to memory
    - Load data in turn from filesystem and move to GPUs
    - Schedule effective data placement



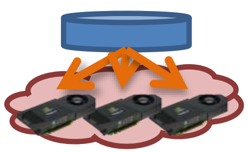


Proposal: Multi-GPU GIM-V with Load Balance Optimization



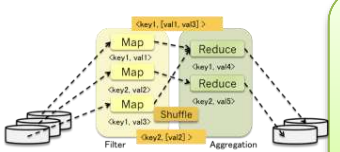
Graph Application

PageRank



Graph Algorithm

Multi-GPU GIM-V



MapReduce Framework

Multi-GPU Mars

Platform

CUDA, MPI

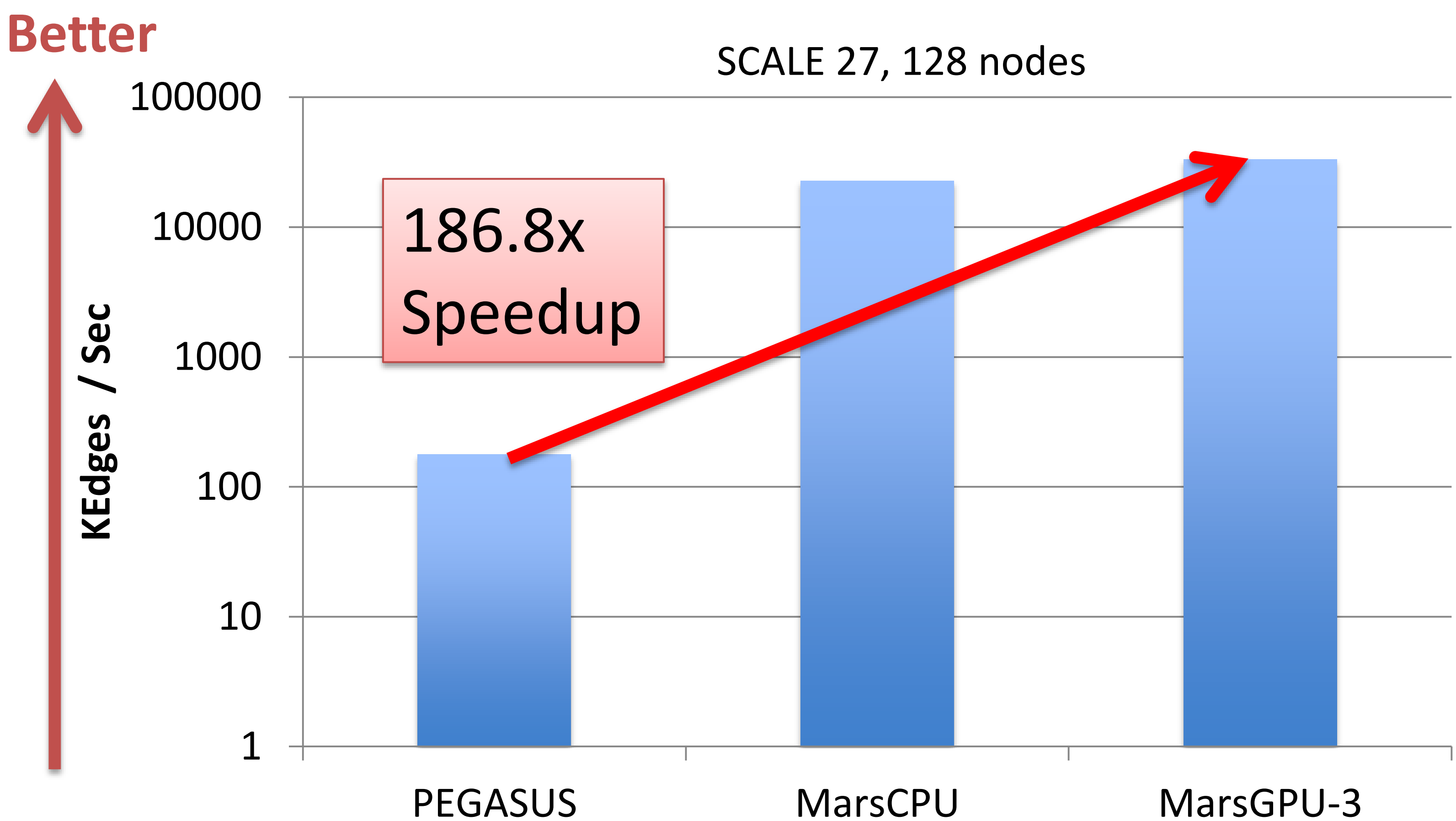
Implement GIM-V on multi-GPUs MapReduce

- Optimization for GIM-V
- Load balance optimization

Extend an existing GPU MapReduce framework (Mars) for multi-GPU

Outperform Hadoop-based Implementation

- PEGASUS: a Hadoop-based GIM-V implementation
  - Hadoop 0.21.0
  - Lustre for underlying Hadoop's file system



(R. Stevens Presenation) NVDIMM

Gradient Machine

- Nodes with various DRAM:NVRAM ratios
  - – 16 GB RAM : 64 GB NVRAM (1:4) – comp node
  - – 16 GB RAM : 256 GB NVRAM (1:16) – hybrid_1 node
  - – 16 GB RAM : 1 TB NVRAM (1:64) – hybrid_2 node
  - – 16 GB RAM : 4 TB NVRAM (1:256) – store node
- Machine consists of sets of nodes of various types (X of comp, Y of store, etc.)
- Supernode could consistent of node collections with dynamic network provisioing

Imagine 1 M nodes of each type..

64 PB DRAM

5540 PB of NVRAM

85x DRAM storage

Jobs run where storage requirements are met

Data can migrate

Compute can migrate

Bandwidth per NVRAM BYTE varies

Bandwidth per DRAM byte is constant

Scaling up to Petabyte/s I/O EBD 2017-18

Process 100 ExaB/Day, 30 ZetaB/Year(松岡素案)

Node
24 DIMMs
384GB DRAM
12TB Flash
600GB/s DRAM BW,
24~48GB Flash BW
20GB/S NW BW
(bidirectional)
6TFlops
480W, $17,000

Rack
4 cabinets/64 nodes
25TB DRAM
786TB Flash
50 TB/s DRAM BW
1.54TB/s Flash BW
1.28TB/s NW BW
384TFLops
30.7KW, $1 mil

Cabinet
16 nodes
6.1TB DRAM
197TB Flash
10TB/s DRAM BW
384GB Flash BW
320GB/s NW BW
96TFLops
7.7KW, $270,000

IDC/SC
650 Racks (~ES)
41,600nodes
16PB DRAM
511PB Flash
25.6PB/s DRAM BW
1PB/s Flash BW
(x1000 K-comp HDD)
250PFlops DFP
500PFlops SFP
830TB/s NW BW
20MW, $700 million

Hardware/Software Approaches

- Hardware support for nv storage on node in memory address space
- Hardware support for variety of operators against storage (hashing, indexing, search, etc.) $\Rightarrow$ CAM
- Language support for data intrinsics
- Support for scripting DSLs bound to high-performance data specific libraries
- Libraries/filters for replacing explicit I/O

Big Interactivity

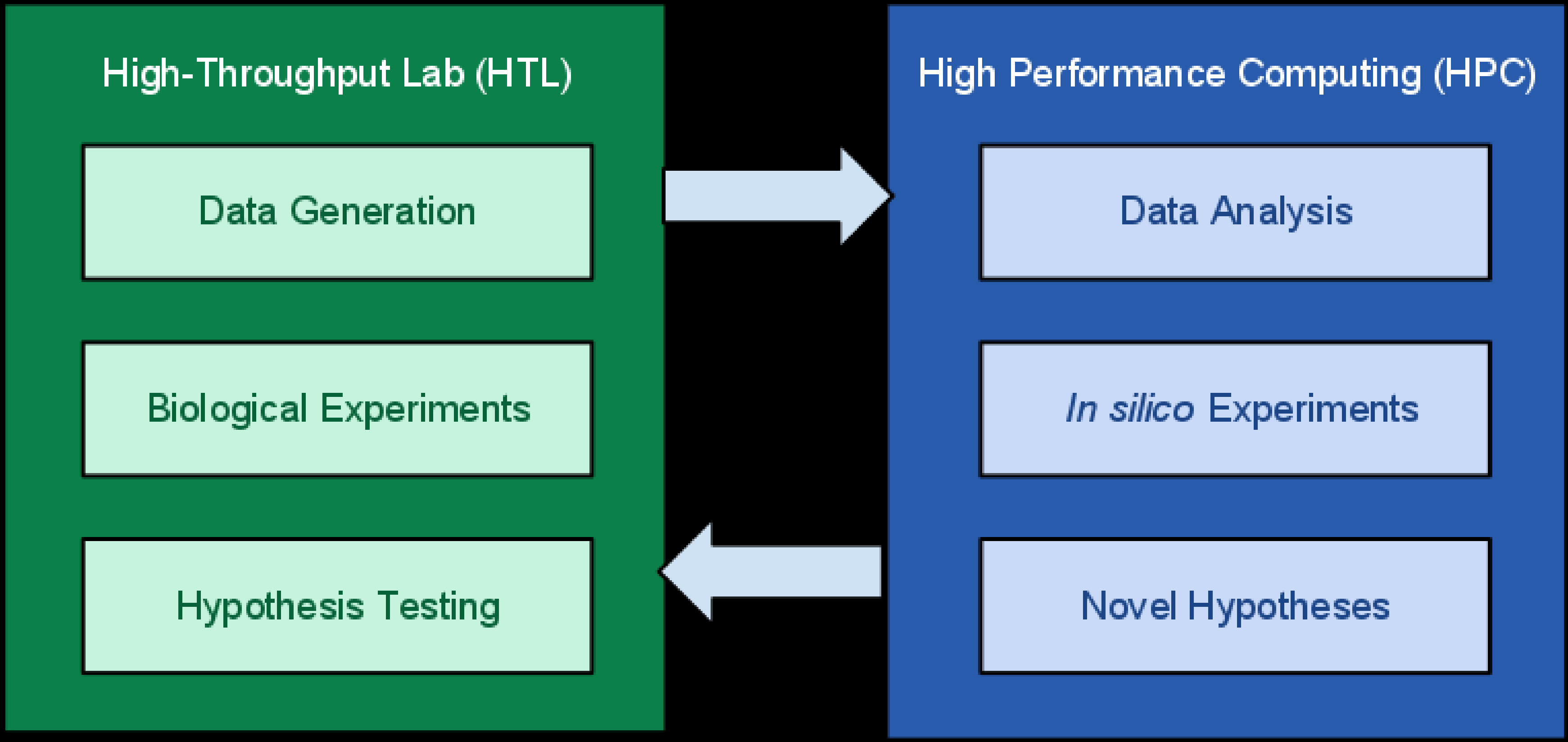
- Support for acquiring multiple I/O nodes with multiple external network connections
- Support for composing connections with outboard rendering engines, etc.
- Flexible input devices (cameras, tracking, audio, etc.)
- Support for jobs proxies in social media, interactive devices, mobile
- Capture and playback support (tutorials)
- Archive and annotate (desktop capture)
- Jobs pause forward and reverse

Research Areas: New Algorithms, Software and Hardware Architectures Needed

- Sequence mapping, assembly, alignment, clustering
- Pattern and feature matching and discovery in complex data models
- Domain specific data compression methods sequence, vector spaces
- Error detection and correction methods in sequence, vector spaces
- Heuristic search over complex data models
- Constraint based methods for fitting, mixed/integer linear programming
- Text indexing, search, query methods

- (alg, hw) Sequence assembly and characterization (hw) Pattern matching architectural support
- (alg, sw) Approximate matching methods for patterns in complex data models
- (sw) Workflow infrastructure for parallel systems and cloud based services
- (sw) Interactive workspaces and rapid prototyping environment with DSLs and in memory database

Automate and Accelerate



Plans & Future

- Building on IESP Success
- NSF Support
- Series of meetings (18 mo)
- Report
- Group picture before lunch