End-User Examples: High Performance Data Analysis (HPDA)

Steve Conway
IDC Research Vice President
HPC & High Performance Data Analysis
sconway@idc.com

July 2014

High Performance Data Analysis

Needs HPC resources

- Complex algorithms
- Near-real time (often)
- Data “long” and “wide”
- On premise or in cloud

Simulation & analytics

- Search, pattern discovery
- Iterative methods
- Established HPC users + new commercial users

Data of all kinds

- The 4 V’s: volume, variety, velocity, value
- Structured, unstructured
- Partitionable, non-partitionable
- Regular, irregular patterns

Different Systems for Different Jobs

IDC
Analyze the Future

Partitionable Work

- Most jobs are here
- **Search** (e.g., Jeopardy Watson)
- Global memory not so important
- Standard clusters + Hadoop, Cassandra, HPCC, etc.

vs.

Non-Partitionable Work

- Toughest jobs (e.g., graphing)
- **Dynamic pattern discovery** (SGI UV, YarcData Urika, medical Watson, et al.)
- Global memory important
- Systems turbo-charged for data movement

HPC Adoption Timeline (Examples)

Ford

Morgan Stanley

P&G

SCHRÖDINGER, INC.

GEICO

TOYOTA

DreamWorks Animation SKG

Walmart

CMS Centers for Medicare & Medicaid Services

PayPal

BOEING

bp

SAMSUNG

MERCK

PING

UNITED STATES POSTAL SERVICE

AIRBUS

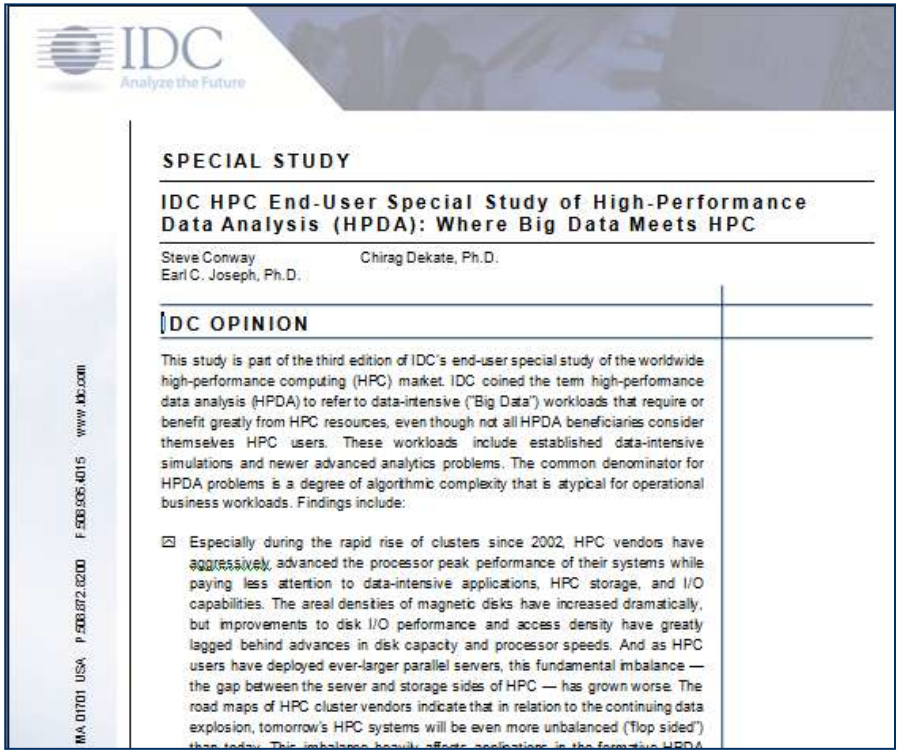
Panasonic

MAYO CLINIC

emcien

1960 1970 1980 1990 2000 2012

IDC 2013 Worldwide HPC End-User Study: HPDA Top Findings



IDC
Analyze the Future

SPECIAL STUDY

IDC HPC End-User Special Study of High-Performance Data Analysis (HPDA): Where Big Data Meets HPC

Steve Conway Chirag Dekate, Ph.D.
Earl C. Joseph, Ph.D.

IDC OPINION

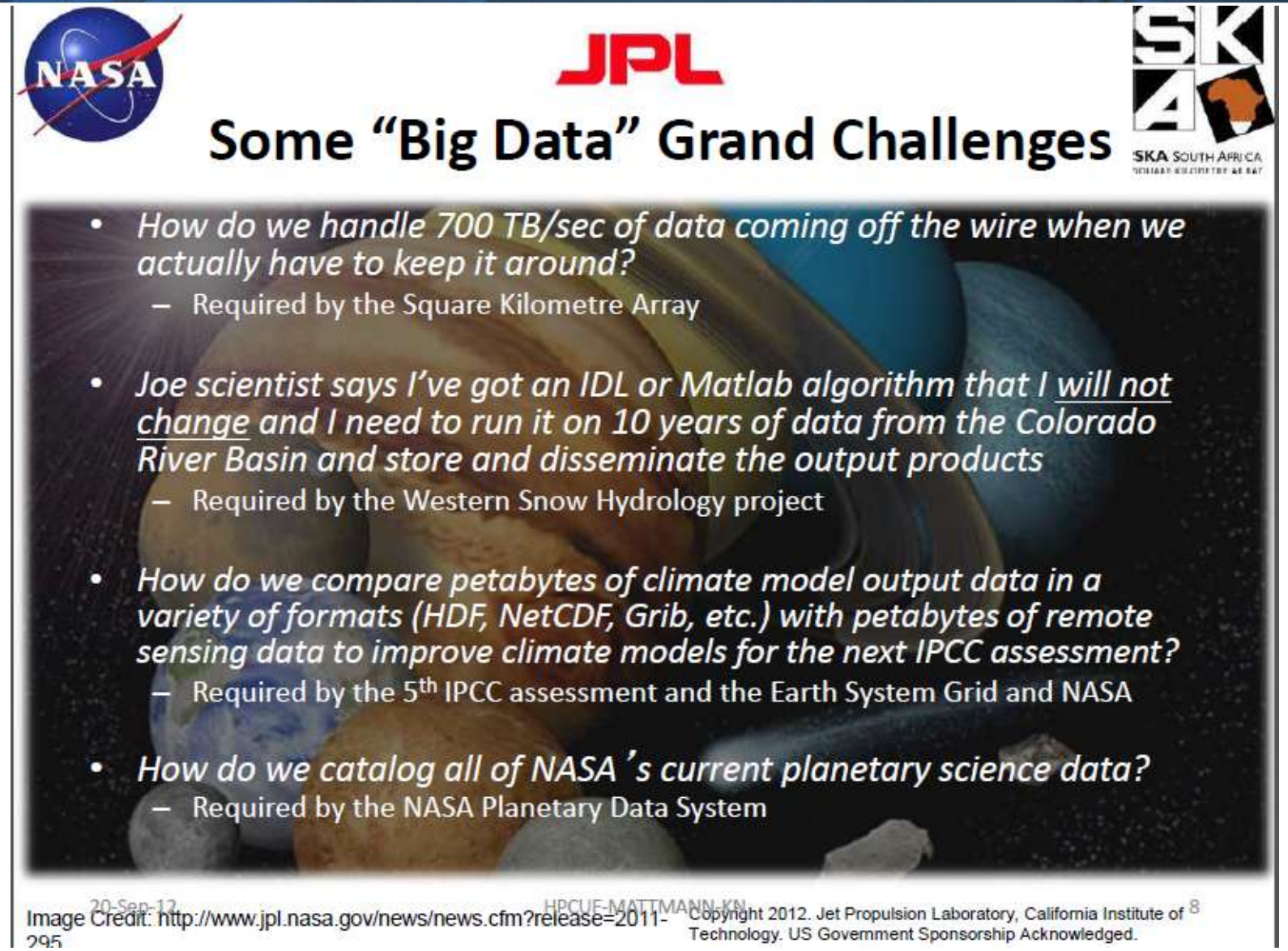
This study is part of the third edition of IDC's end-user special study of the worldwide high-performance computing (HPC) market. IDC coined the term high-performance data analysis (HPDA) to refer to data-intensive ("Big Data") workloads that require or benefit greatly from HPC resources, even though not all HPDA beneficiaries consider themselves HPC users. These workloads include established data-intensive simulations and newer advanced analytics problems. The common denominator for HPDA problems is a degree of algorithmic complexity that is atypical for operational business workloads. Findings include:

- Especially during the rapid rise of clusters since 2002, HPC vendors have aggressively advanced the processor peak performance of their systems while paying less attention to data-intensive applications, HPC storage, and I/O capabilities. The areal densities of magnetic disks have increased dramatically, but improvements to disk I/O performance and access density have greatly lagged behind advances in disk capacity and processor speeds. And as HPC users have deployed ever-larger parallel servers, this fundamental imbalance — the gap between the server and storage sides of HPC — has grown worse. The road maps of HPC cluster vendors indicate that in relation to the continuing data explosion, tomorrow's HPC systems will be even more unbalanced ("flop sided") than today. This imbalance heavily affects applications in the formative HPDA

MA 01701 USA P 508.872.8200 F 508.935.4015 www.idc.com

- 67% of the sites perform HPDA work (data-intensive simulation and/or advanced analytics).
- On average, HPDA consumes 30% of compute cycles.
- 29% of sites use Hadoop
- Major pain points worth 10-15% premium pricing:
  - Interconnects between nodes
  - External I/O and storage

Big Science: Big Data Challenges Are Growing…



Total Revenue Protection Program

- Processing Requirements
  - Rate
    - 4 billion mail scans per day peak (74,000 per second)
  - Geographic Scope
    - Incoming mail from 275 Processing and Distribution Centers
    - Outgoing mail to 33,000 postal operated facilities
  - Objective
    - To find, track and reject mail pieces due to:
      - Duplicate postage
      - Short Pay
      - Ineligible Discounts

FedCentric TECHNOLOGIES

U.S. Postal Service

Enterprise Supercomputing

Why Real Time Fraud Detection?

Save time... print your postage online.

Print exact postage for letters and packages using just your PC and printer.

Print Postage Stamps

- Print any denomination
- Use for letters or packages
- Never run out of stamps again

DETAILS

Stamps.com... Your own personal Post Office open 24 hours a day.

Developed in conjunction with the United States Postal Service,℠ Stamps.com is a revolutionary software-based service that allows you to calculate and print official USPS postage right from your PC.

NO ADDITIONAL HARDWARE REQUIRED. Stamps.com even keeps track of all your postal spending using your client codes, and can even recommend optimal delivery methods, formats and more. Plus, Stamps.com gives you postage discounts you can't even get at the Post Office or with a postage meter.

5

Fed Centric TECHNOLOGIES

U.S. Postal Service

TRP Results using MCDB &TimesTen

Pre-MCDB

1. 509 row inserts per second (RIPS)
2. Direct path load option a partial solution (2000 RIPS)
3. 275 Million Transactions per 15 hour processing window created backlog during peak processing windows
4. Revenue Protection performed as a batch data warehouse process, run 3 – 12 hours after Mailpiece scan

With MCDB Deployed

1. 190,222 RIPS (3 Threads)
2. 1,091,018 RIPS (18 Threads)
3. Processed 4 B Transactions in less than 6 hours
4. Revenue Protection is performed in real-time upon first scan

MCDB = memory-centric database

10

CMS: Government Health Care Fraud

- 5 separate databases for the big USG health care programs under Centers for Medicare and Medicaid Services (CMS)
- Estimated fraud: $150B-$450B. <$5B caught today)
- ORNL, SDSC have evaluation contracts to unify the databases and perform fraud detection on various architectures.

Use Case: PayPal

Fraud Detection / Internet Commerce

Slides and permission provided by PayPal, an eBay company

The Problem

Detecting fraud in 'real time' as millions of transactions are processed between disparate systems at volume.

Finding suspicious patterns that we don't even know exist in related data sets.

Ability to create and deploy new fraud models into event flows quickly and with minimal effort.

Provide environment for fraud modeling, analytics, visualization, M/R, dimensioning and further processing.

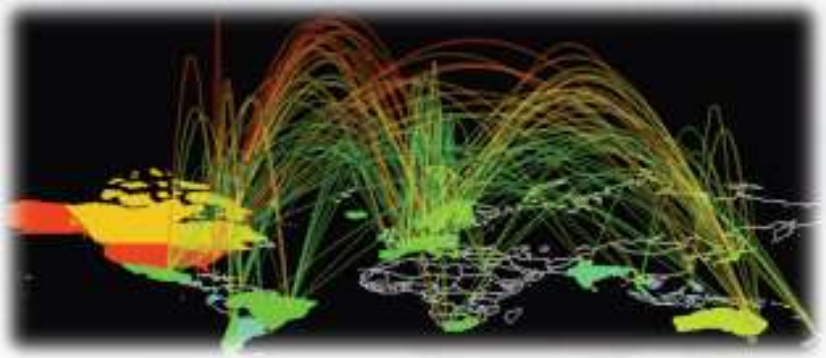
PayPal: Next Uses for HPC

- After this success, PayPal now plans further uses for HPC:
  - Managing the whole PayPal IT infrastructure
  - Affinity marketing to consumers ("Beacon" project)
- Parent company eBay is not using HPC yet.

"Clearly understand that HPC is not a mass consumption technology where we enable everyone in our organization with it. This is a deep engineering function. It's custom built and includes writing software to solve cutting-edge problems ... Think of HPC not as an IT function but as a competitive business advantage. There's a hard link between HPC and PayPal's top line and bottom line."

PayPal CTO Jim Barrese (IDC interview, 2013)

Use Case: Network Security



Network Intrusion Detection

32,000 Items (src & dest. IP address,ports, days, times, activities, etc.)

2.57 million Transactions

No rules or queries required. Auto-detect intrusion patterns and surface suspicious activity.

1	Src-10.2.197.245 →	Start_hour:9	Thu	Dst-80	6,806	Strong	100%
		Dst-154.241.88.201					
11	Src-10.2.197.245 →	Priority:1	Dst-80	Dst-154.241.88.201	6,804	Strong	100%

emcien
Converting Data into Value

Schrödinger: Cloud-based Lead Discovery for Drug Design

IDC
Analyze the Future

Metric from March, 2012	Count
Compute Hours of Work	109,927 hours
Compute Years of Work	12.55 years
Total # Cores/Servers	51132 cores, 6742 servers
Infrastructure Value	~ $20,000,000
AWS Regions	All (7: us-east, us-west1, us-west2, eu-west, sa-east, ap-northeast, ap-southeast)

Using CycleCloud & AWS:
Impossible run in 3 hours for $4,828/hr
Today's pricing < $1,000/hr

CYCLECOMPUTING

Outcomes-Based Medical Diagnosis and Treatment Planning

- Enter the patient's history and symptomology.
- While patient is still in the office, sift through millions of archived patient records for relevant outcomes.
- Provider considers the efficacies of various treatments for "similar" patients (but is not bound by the findings).
- Ergo, this functions as a powerful decision-support tool.
- Benefits: better outcomes + rein in costly outlier practices

Optum Labs: UHG-led Collaborative to Advance Big Data in Health Care

- $500 million center planned in Cambridge, MA
- Pre-competitive, open research
- Contributors sit on governance board (e.g., Mayo)
- Long-term goal: enable outcomes-based medicine

IDC HPDA Server Forecast

- Fast growth from a small starting point: $1.2B (€900M by 2016)
- HPDA ecosystem >$2B (€1.5B) in 2016

TABLE 2

IDC Worldwide High Performance Data Analysis (HPDA) Server Revenues

($ Millions)

	2009	2010	2011	2012	2013	2014	2015	2016	CAGR '11-'16
WW HPC Server Sales	8,637	9,498	10,300	11,098	11,397	12,371	13,485	14,621	7.3%
WW HPDA Server Sales	535	603	673	744	786	881	1,109	1,253	13.3%
HPDA Portion	6.2%	6.3%	6.5%	6.7%	6.9%	7.1%	8.2%	8.6%	5.6%

Source: IDC 2013

IDC HPDA Storage Forecast

- Storage is the fastest-growing HPC market (8.4% CAGR, 2011-16) and HPDA storage will grow even faster (18.1% CAGR).

TABLE 2

Worldwide High-Performance Data Analysis Storage Revenue, 2009–2016 ($M)

	2009	2010	2011	2012	2013	2014	2015	2016	2011–2016 CAGR (%)
HPC storage	3,023.0	3,325.9	3,761.5	4,194.0	4,349.8	4,739.1	5,163.2	5,625.3	8.4
Share as total HPC server revenue (%)	35.0	35.0	36.5	37.8	38.2	38.3	38.3	38.5	1.0
HPDA storage	262.2	301.5	343.0	387.0	432.2	519.9	676.5	789.5	18.1
Big Data attach rate (%)	49.0	50.0	51.0	52.0	55.0	59.0	61.0	63.0	4.3

Source: IDC, 2013

Summary: HPDA Market Opportunity

- **HPDA: simulation + newer high-performance analytics**
  - IDC predicts fast growth from a small starting point
- **HPC and high-end commercial analytics are converging.**
  - Algorithmic complexity is the common denominator
- **Economically important use cases are emerging**
  - Which ones will become attractive markets?
- **No single HPC solution is best for all problems**.
  - Clusters with MR/Hadoop will handle most but not all work (e.g., graph analysis)
- **IDC believes our growth estimates could be conservative**.

Questions?

Please email:
hpc@idc.com

Check out:
www.hpcuserforum.com
www.hpcuserforum.com/ROI
www.idc.com