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

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

April 2014

High Performance Data Analysis

Needs HPC resources

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

Data of all kinds

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

Simulation & analytics

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

Different Systems for Different Jobs

Partitionable Work

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

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

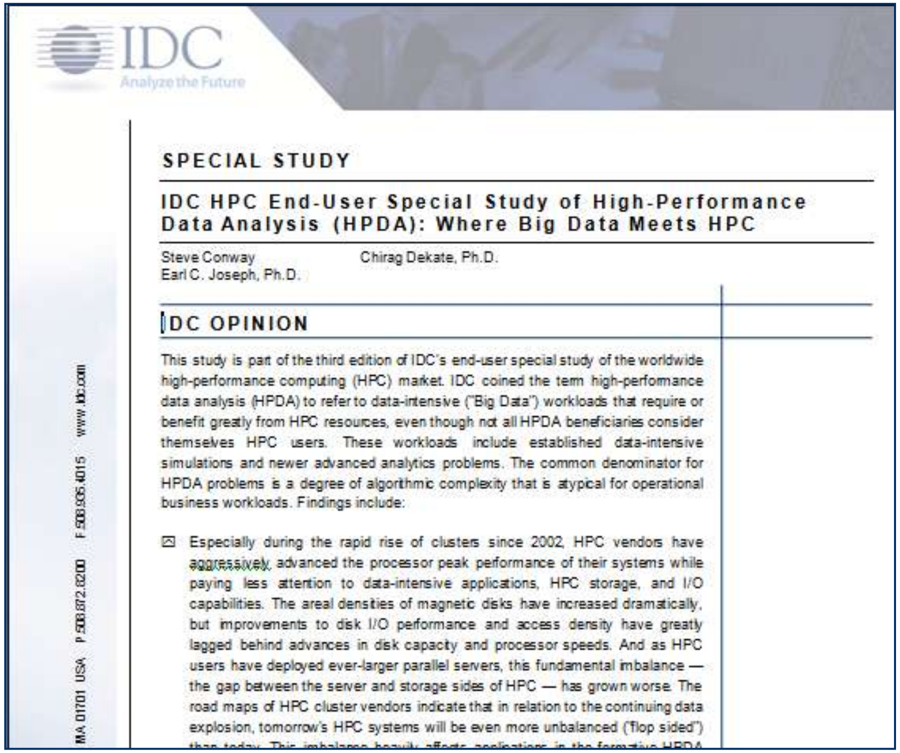
vs.

HPC Adoption Timeline (Examples)

Ford
Morgan Stanley
P&G
SCHRÖDINGER, INC.
GEICO
TOYOTA
DREAMWORKS ANIMATION SKG
Walmart
CMS CENTERS for MEDICARE & MEDICAID SERVICES
BOEING
bp
SAMSUNG
PayPal
VISA
MasterCard
DISCOVER
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
Earl C. Joseph, Ph.D.
Chirag Dekate, Ph.D.

IDC OPINION

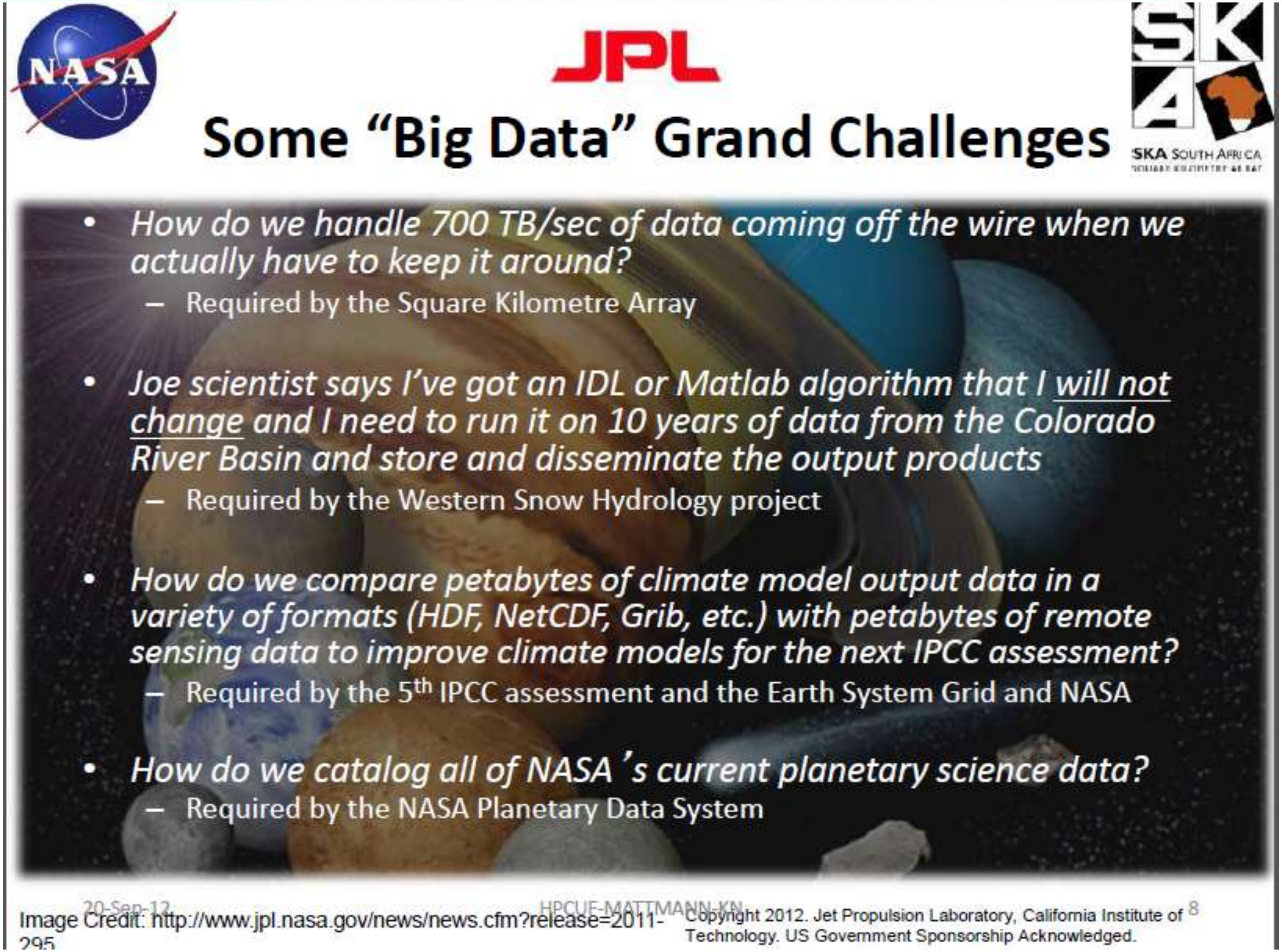
This study is part of the third edition of IDC's end-user special study of the worldwide high-performance computing (HPC) market. IDC coined the term high-performance data analysis (HPDA) to refer to data-intensive ("Big Data") workloads that require or benefit greatly from HPC resources, even though 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…



Some "Big Data" Grand Challenges

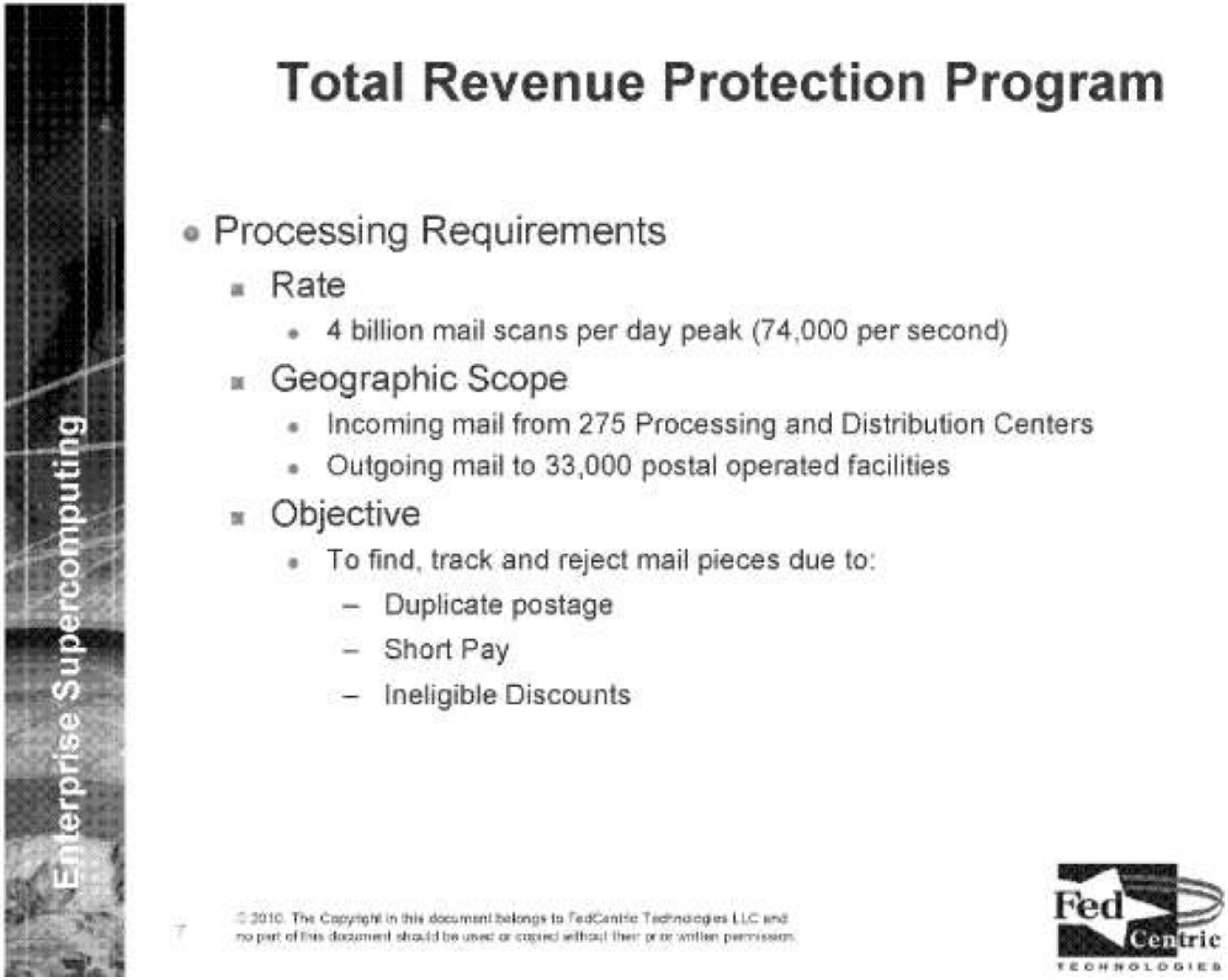
- *How do we handle 700 TB/sec of data coming off the wire when we actually have to keep it around?*
  - Required by the Square Kilometre Array
- *Joe scientist says I've got an IDL or Matlab algorithm that I will not change and I need to run it on 10 years of data from the Colorado River Basin and store and disseminate the output products*
  - Required by the Western Snow Hydrology project
- *How do we compare petabytes of climate model output data in a variety of formats (HDF, NetCDF, Grib, etc.) with petabytes of remote sensing data to improve climate models for the next IPCC assessment?*
  - Required by the 5th IPCC assessment and the Earth System Grid and NASA
- *How do we catalog all of NASA's current planetary science data?*
  - Required by the NASA Planetary Data System

20-Sep-12 HPCUF-MATTMANN-KN 8

Image Credit: http://www.jpl.nasa.gov/news/news.cfm?release=2011-295

Copyright 2012. Jet Propulsion Laboratory, California Institute of Technology. US Government Sponsorship Acknowledged.

U.S. Postal Service



Enterprise Supercomputing

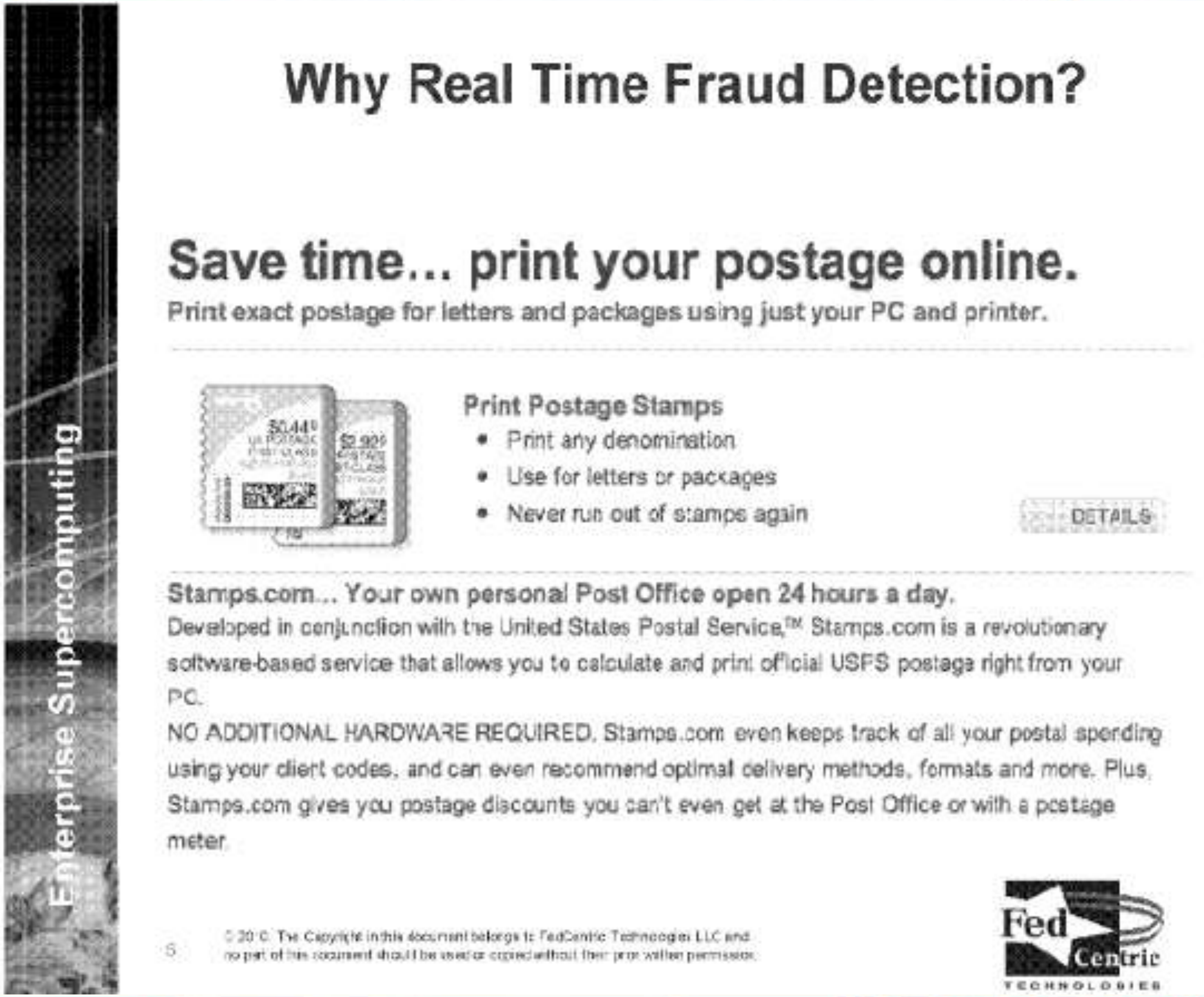
Total Revenue Protection Program

- 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

7

FedCentric TECHNOLOGIES

U.S. Postal Service


Enterprise Supercomputing
$0.44
$2.92
DETAILS
Fed Centric
TECHNOLOGIES

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

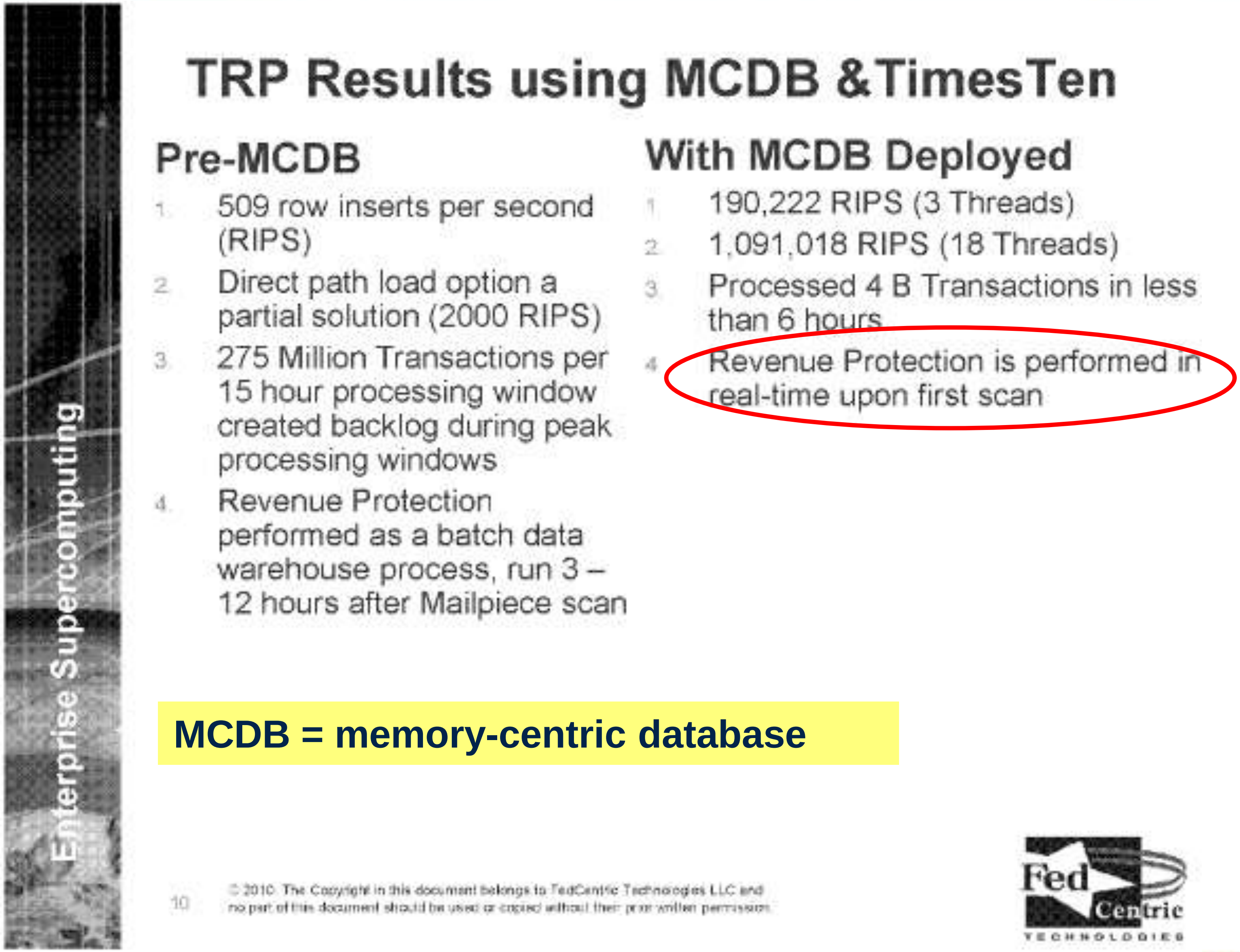
Stamps.com… Your own personal Post Office open 24 hours a day.

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

U.S. Postal Service



Enterprise Supercomputing

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

CMS: Government Health Care Fraud

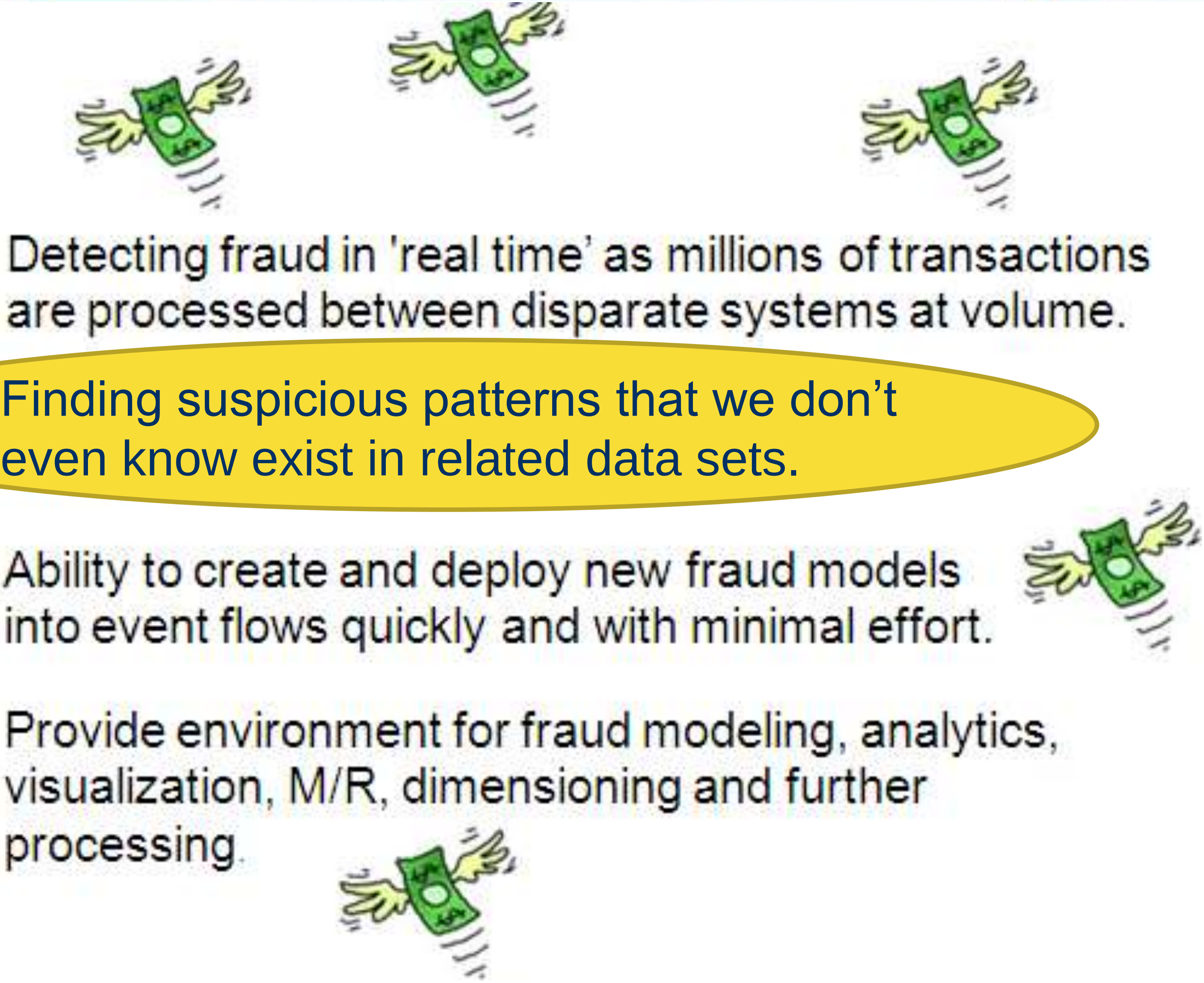
- 5 separate databases for the big USG health care programs under Centers for Medicare and Medicaid Services (CMS)
- Estimated fraud: $150B-$450B. <$5B caught today)
- ORNL, SDSC have evaluation contracts to unify the databases and perform fraud detection on various architectures.

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

IDC Analyze the Future

- 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.

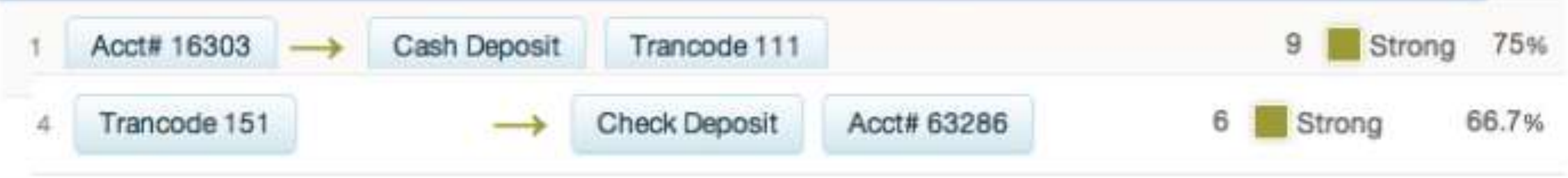
“Clearly understand that HPC is not a mass consumption technology where we enable everyone in our organization with it. This is a deep engineering function. It's custom built and includes writing software to solve cutting-edge problems ... Think of HPC not as an IT function but as a competitive business 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: Banking

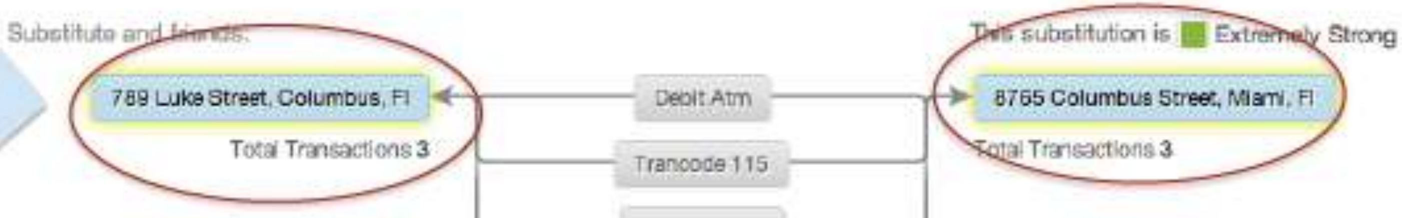
Fraud Detection
480,000 Items (accounts, codes, locations, etc.)
2.94 million Transactions

Revealing top suspicious transaction patterns (account numbers, transaction types,..)



1	Acct# 16303 →	Cash Deposit	Trancode 111	9	🟩	Strong	75%
4	Trancode 151 →	Check Deposit	Acct# 63286	6	🟩	Strong	66.7%

Revealing ID fraud with fuzzy match

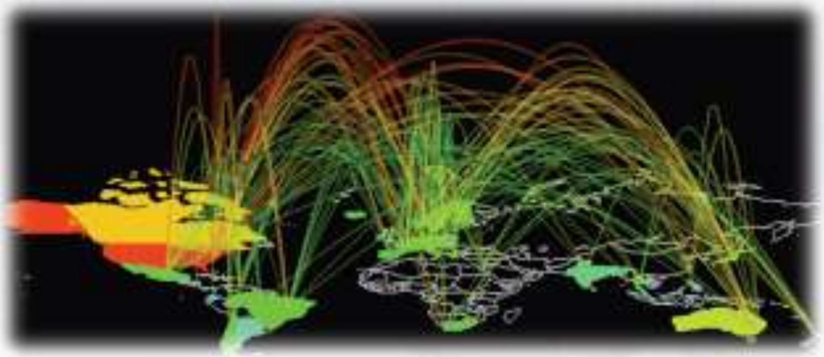


Substitute and friends:
This substitution is 🟩 Extremely Strong
789 Luke Street, Columbus, Fl
Total Transactions 3
Debit Atm
Trancode 115
Acct# 64127
8765 Columbus Street, Miami, Fl
Total Transactions 3

← **Friends**
The gray items are "Friends". Each substitute is found in transactions with these friends BUT not with each other.

emcien
Converting Data Into Value

Use Case: Network Security

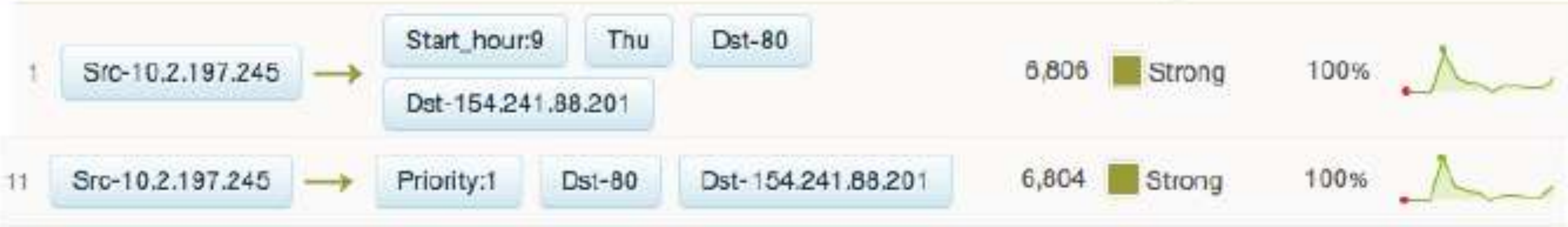


Network Intrusion Detection

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.

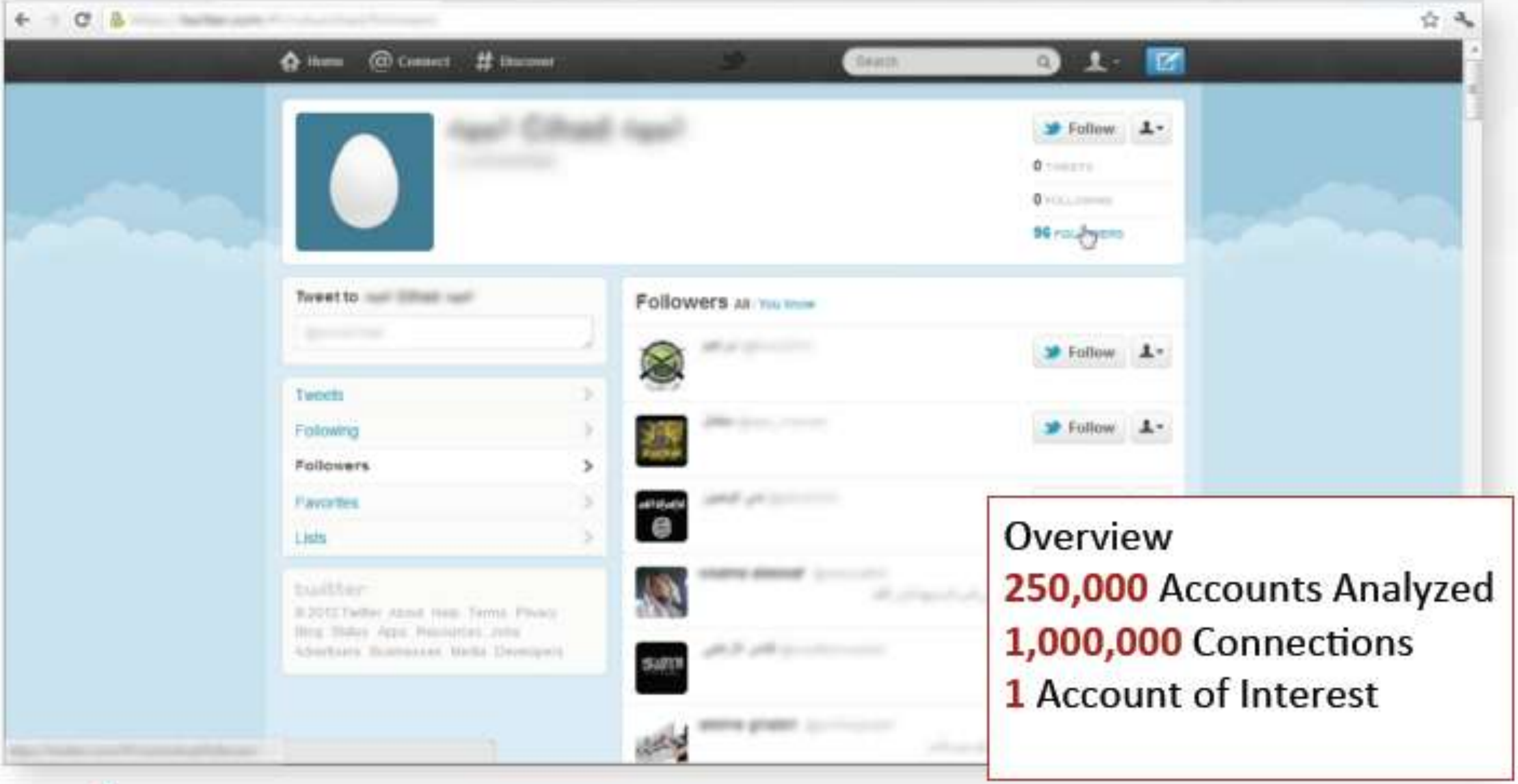


1	Src-10.2.197.245	→	Start_hour:9 Thu Dst-80 Dst-154.241.88.201	6,806	🟩 Strong	100%
11	Src-10.2.197.245	→	Priority:1 Dst-80 Dst-154.241.88.201	6,804	🟩 Strong	100%

emcien
Converting Data Into Value

Use Case: Surfacing Sleeper Cell for Intel

The silent signal – Automatically detecting a sleeper cell



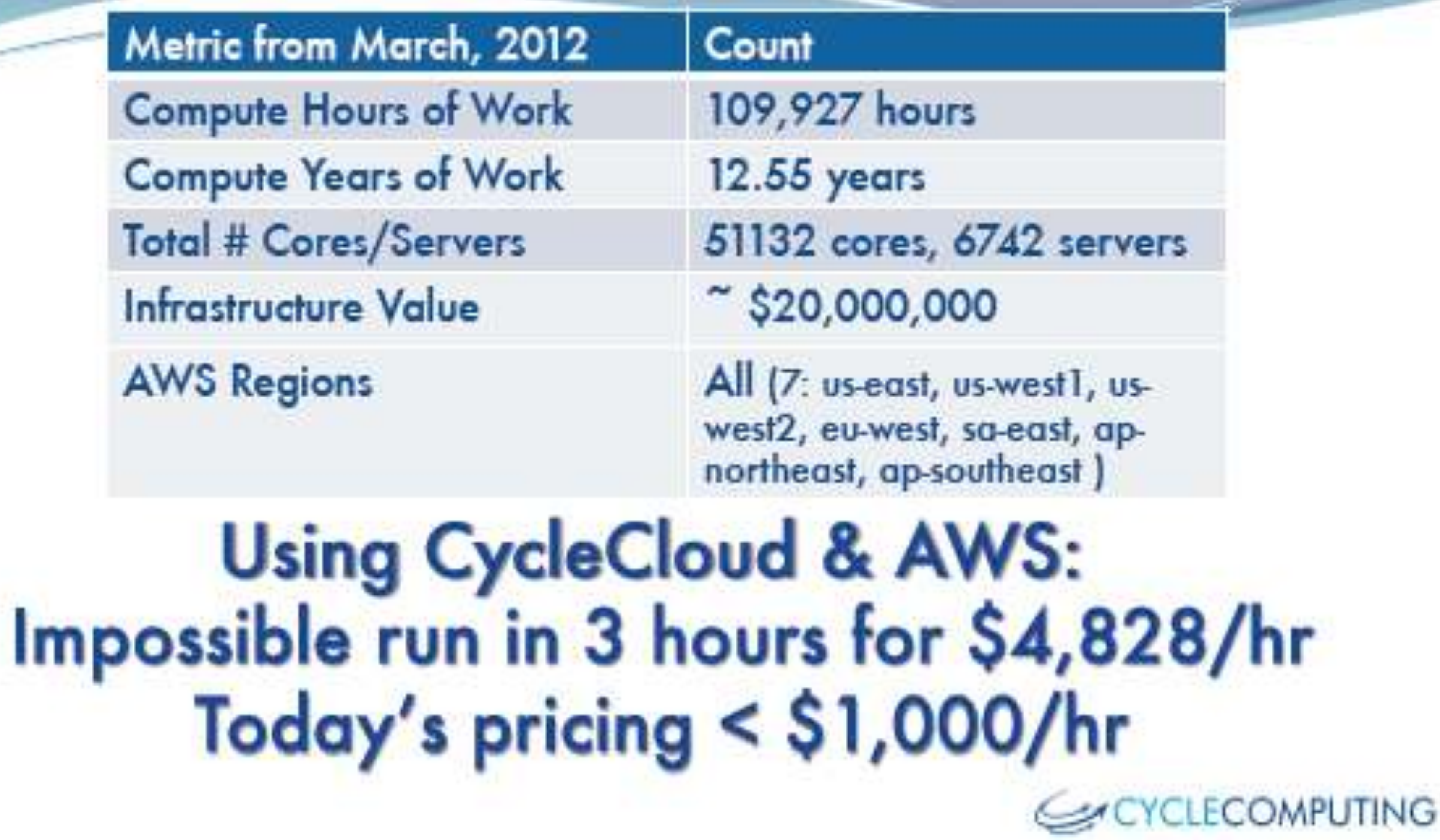
Overview

250,000 Accounts Analyzed

1,000,000 Connections

1 Account of Interest

Schrödinger: Cloud-based Lead Discovery for Drug Design

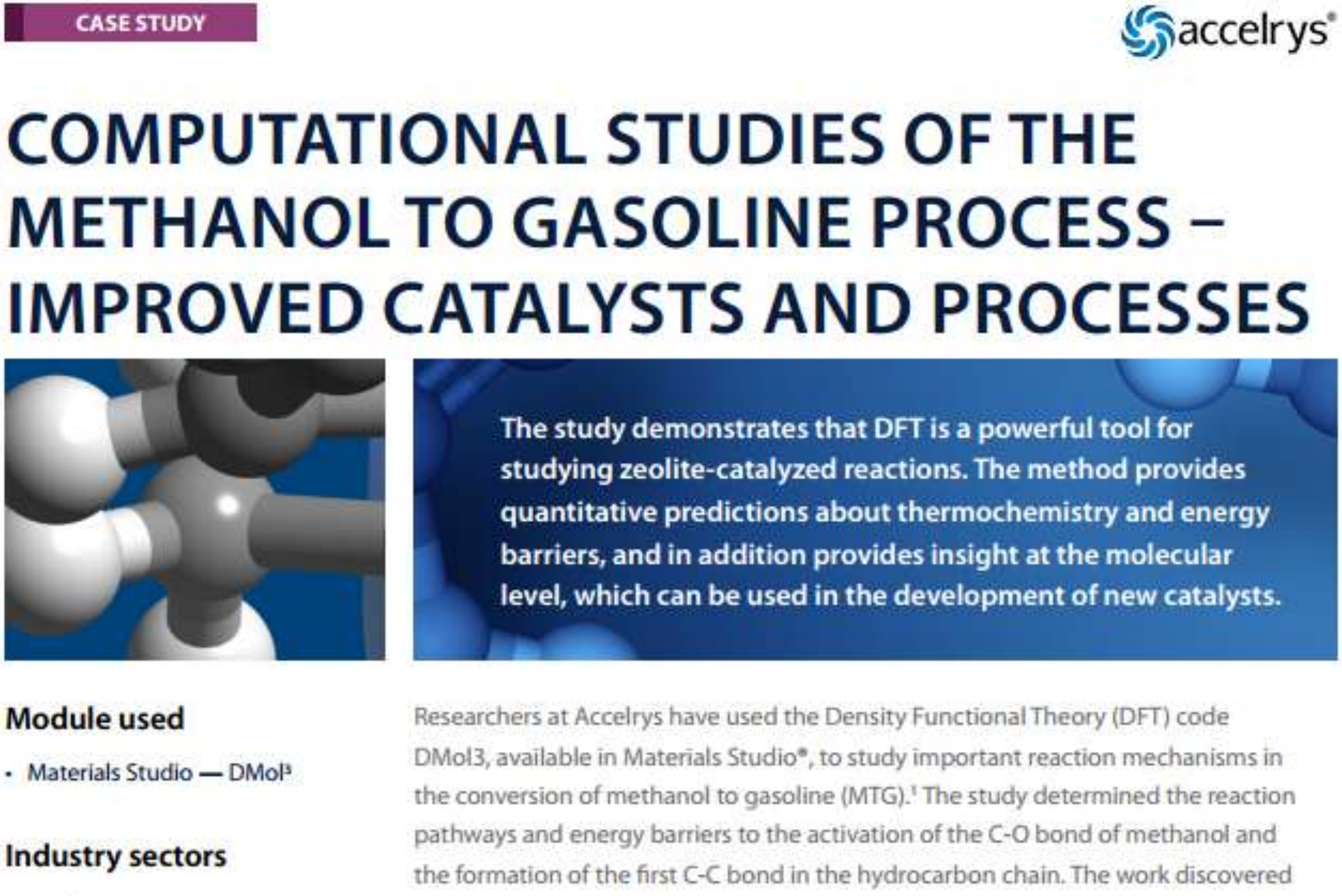


Metric from March, 2012	Count
Compute Hours of Work	109,927 hours
Compute Years of Work	12.55 years
Total # Cores/Servers	51132 cores, 6742 servers
Infrastructure Value	~ $20,000,000
AWS Regions	All (7: us-east, us-west1, us-west2, eu-west, sa-east, ap-northeast, ap-southeast)

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

CYCLECOMPUTING

Accelrys Uses HPDA for Drug Discovery and Materials Science



CASE STUDY

COMPUTATIONAL STUDIES OF THE METHANOL TO GASOLINE PROCESS – IMPROVED CATALYSTS AND PROCESSES

> The study demonstrates that DFT is a powerful tool for studying zeolite-catalyzed reactions. The method provides quantitative predictions about thermochemistry and energy barriers, and in addition provides insight at the molecular level, which can be used in the development of new catalysts.

Module used

- Materials Studio — DMol^3

Industry sectors

Researchers at Accelrys have used the Density Functional Theory (DFT) code DMol3, available in Materials Studio®, to study important reaction mechanisms in the conversion of methanol to gasoline (MTG).[1] The study determined the reaction pathways and energy barriers to the activation of the C-O bond of methanol and the formation of the first C-C bond in the hydrocarbon chain. The work discovered

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

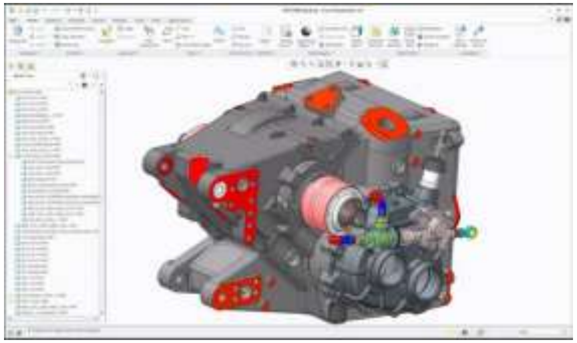
- $500 million center planned in Cambridge, MA
- Pre-competitive, open research
- Contributors sit on governance board (e.g., Mayo)
- Long-term goal: enable outcomes-based medicine

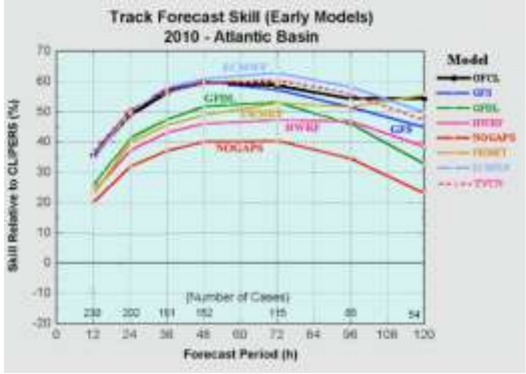
OPTUM
Good for the system.™

MAYO
CLINIC

Iterative Methods (Cumulative Data)

- Parametric modeling (product design)
- Stochastic modeling (financial)
- Ensemble modeling (weather/climate)





IDC HPDA Server Forecast

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

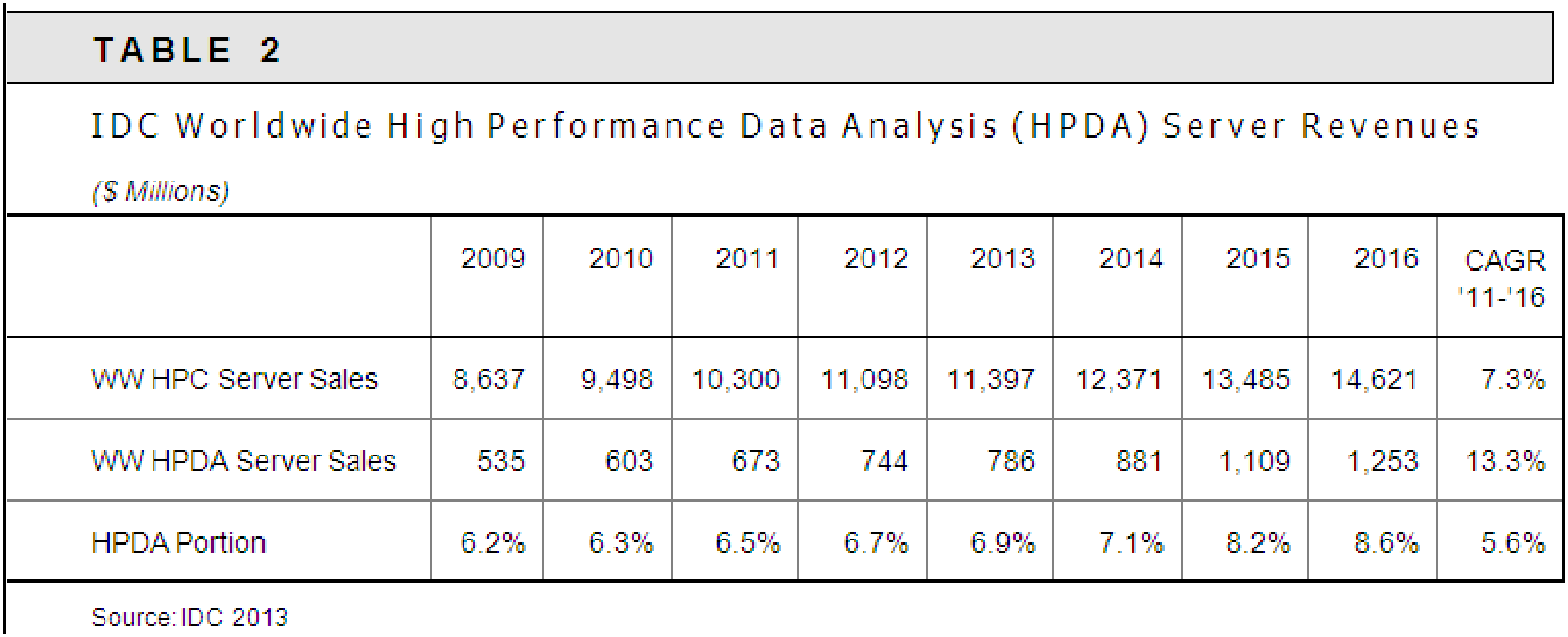


TABLE 2

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).

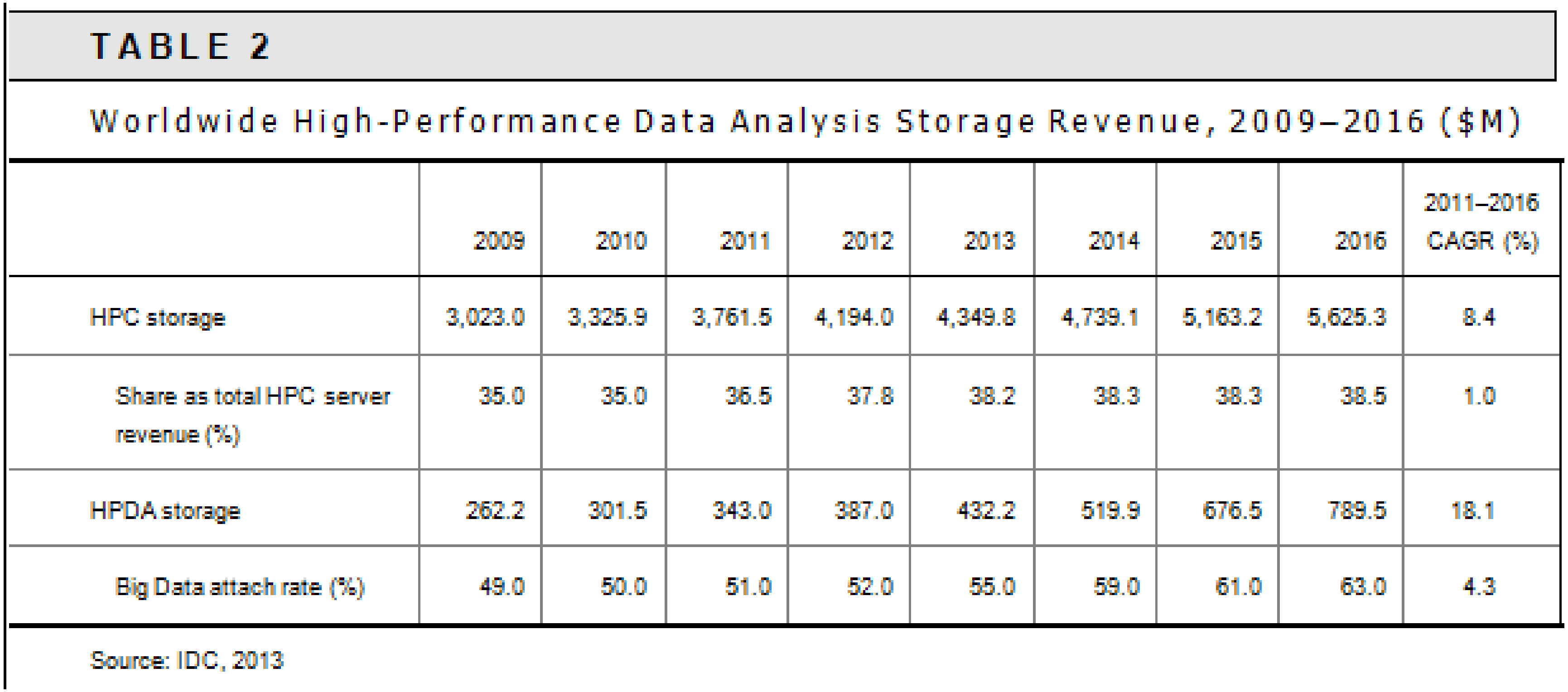


TABLE 2

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