CIO/NII
Enabling Net-Centric Operations

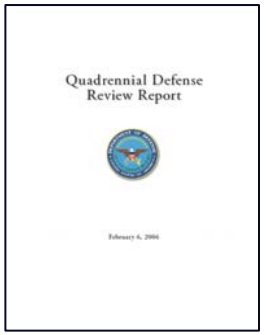
Accessing and Using Web-Scale Data Web 3.0 Style

Contacts:
Mike Moore (john.moore@osd.mil)
703-602-0943
Office of the DoD CIO

Changing Strategic Environment

- **Challenge – UNCERTAINTY**
  - **"Transforming … to an expeditionary force, providing greater flexibility to contend with uncertainty in a changing strategic environment."**

- **Response – AGILITY**
  - **"We must develop a mix of agile and flexible capabilities to mitigate uncertainty."**



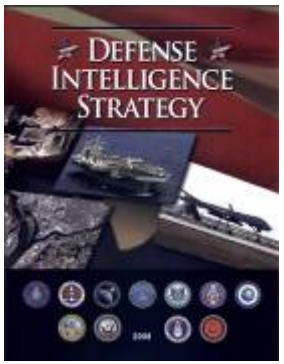
- Enterprise-wide: Battlefield Applications; Defense Operations; Intelligence Functions; Business Processes

- Emphasis Shift: From moving the user to the data – to moving data to the user

Confronting Uncertainty with Agility

Changing Information Needs

- Targets entail complex threat networks
  - Needles in needle stacks
  - Multiple locations crossing organizational boundaries
  - Spontaneous with short duration
  - On-the-move
- It's about the data – all the data
  - Multiple sources and disciplines
  - Turning data mountains into knowledge – at the atomic level
  - Global access for all organizations
- Expertise has moved to the edge
  - Multiple locations
  - Multiple perspectives



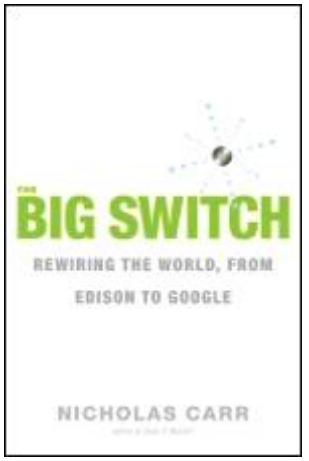
"Extend the full advantage of the U.S. intelligence enterprise to all defense users to ensure timely and accurate decisions."

IT at a Crossroads

- Increased Parallelism
  - New Moore's Law - 2X processors per chip generation
  - Parallel software industries emerging to address challenges
  - Redundant networks and storage increasing performance
- Increased Virtualization
  - Processing, Storage, Bandwidth, Delivery
- Everything as a Service in the Clouds
  - Bell's Law - Every decade new, lower priced computer class forms with new programming platform, network, and interface resulting in new usage and industry.
  - Platform, Software, Data, Expertise
- Increased Outsourcing of Core Elements
  - "By 2012, 80 percent of Fortune 1000 companies will pay for some cloud computing service, and 30 percent of them will pay for cloud computing infrastructure." *Gartner*

The World is Moving to the Clouds

What are Clouds?



- **IBM:** "A cloud is an IT service delivered to users that has:
  - A user interface that makes the infrastructure underlying the service transparent to the user
  - Reduced incremental management costs when additional IT resources are added
  - Services oriented management architecture
  - Massive Scalability"

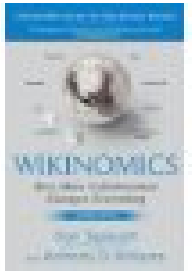
- **Forrester:** "An abstracted, fabric-based infrastructure that enables dynamic movement, growth, and protection of services that is billed like a utility. … cloud computing is looking like a classic disruptive technology"

- **Gartner:** "Cloud computing is a style of computing where massively scalable IT-related capabilities are provided 'as a service' across the Internet to multiple external customers"

- **The 451 Group:** "The cloud is IT, presented as a service to the user, delivered by virtualized resources that are independent of location."

Clouds Bring Enterprise Power to Users

Web 3.0?



- Web 1.0 – Everyone Can Transact
  - E-Bay, Amazon, Search, E-mail, etc
- Web 2.0 – Everyone Can Participate
  - MySpace, Facebook, Flickr, YouTube, del.icio.us, AIM, Twitter, Wikipedia, Skype, etc
- Web 3.0 – Everyone Can Innovate
  - Google Docs, CUBiT, Amazon Web Services, iGoogle, AdSense, Salesforce.com, CODA, InnoCentive, SuccessFactors, etc

My Other Computer is a Data Center!

See: http://www.techcrunchit.com/2008/08/01/welcome-to-web-30-now-your-other-computer-is-a-data-center/

Contrary Views

- Larry Ellison, Oracle CEO at Oracle OpenWorld, September 2008 – "The interesting thing about cloud computing is that we've redefined cloud computing to include everything that we already do. I can't think of anything that isn't cloud computing with all of these announcements. … What is it? It's complete gibberish. It's insane. When is this idiocy going to stop? … We'll make cloud computing announcements. I'm not going to fight this thing. But I don't understand what we would do differently in the light of cloud."[1]

- Richard Stallman, founder of the Free Software Foundation to The Guardian, September 2008 – "One reason you should not use Web applications to do your computing is that you lose control. It's just as bad as using a proprietary program. … It's stupidity. It's worse than stupidity: it's a marketing hype campaign."

[1] See - http://blogs.wsj.com/biztech/2008/09/25/larry-ellisons-brilliant-anti-cloud-computing-rant/

Where are the Clouds?

Sample Cloud Users

Flickr
SmugMug
FOX Interactive Media
PowerSet
Yahoo
MySpace
Facebook
AutoDesk
SalesForce.com
Google
Amazon
JungleDisk
Animoto
Intridea
Enomaly
Heroku
Hyperic

Sample Cloud Providers

DISA
IBM
HP
Skytap
IBM
Elastra
LongJump
XCalibre
Engine Yard
CohesiveFT
Appirio
Cassatt
layeredtech
Kaavo
Oracle
Parascale
Rightscale
Mosso
Sun
Appistry
Microsoft
10gen
3tera
Coghead
GigaSpaces
Limelight
GoGrid
Akamai
EMC
Dell
Joyent
AppNexus
AT&T
Terremark
ClusterResources

User Desktops

"Cloud computing" takes hold as 69% of all internet users have either stored data online or used a web-based software application; 87% of internet users aged 18-29 have done so

Source - September 2008, PEW Internet and American Life Project Report

Switch to Clouds Reaching a Tipping Point?

CIO/NII
Enabling Net-Centric Operations

Open Source Cloud Technologies

- SnapLogic's **Data Integration Framework** – Data pipeline processing tool for SaaS-based environments
- Enomaly's **Enomalism** – Cloud management and provisioning
- UCSB's **Eucalyptus** – Elastic computing management
- Apache/Yahoo!'s **Hadoop** – Parallel, distributed file system, processing engine and multi-dimensional database
- Zvents' **Hypertable** – Parallel, distributed multi-dimensional database
- Intel/CMU/Yahoo!'s **Tashi** – Cloud-based cluster management with integrated compute and storage management for large data problems

Yahoo recently demonstrated a 4,000 commodity node Hadoop Cloud
With 30K+ cores, 16 PB disk storage, and 3.3 TB RAM.
Hadoop accounts for ~30% of AWS usage.

Start from Web-Scaled, Extensible Solutions

A Commercial Example:

washingtonpost.com

amazon web services™

- Problem
  - 17K pages of Hillary Clinton's White House schedule published as unsearchable PDF
  - Convert images into searchable text and deliver them within the same day news cycle
  - Insufficient internal IT resources
- Solution
  - Host user-provided OCR tools on Amazon Elastic Compute Cloud (EC2) Service
  - Acquire needed compute time using Amazon's pay-as-you-go pricing
- Result
  - Conversion completed in 9 Hours on 200 virtual processors
  - Used 1,407 hours of virtual machine time for a final expense of $144.62

Changing Expectations: Agility at a Savings

See: http://aws.amazon.com/solutions/case-studies/washington-post/

Commercial Web-Scale Problems

	Types of Large Data Elements			Driving Requirements		
Problem	**Large Files**	**Large # Files**	**Large Streams**	**Transport**	**Processing**	**Storage**
Network Analysis - 1+ TB/day IM & 240M nodes/month		X			X	X
Retail Mgmt Logistics - 267M items/day & 4+PB warehouse		X			X	X
3D Real-World Simulation - 10^{12} Building Components/Country		X			X	X
Machine Translation - 100B Model Elements w/ 1M Lookups/s		X			X	X
Photo Mgmt - 100+ PB photos/yr		X			X	X
Web Search - 1+ PB of Web pages; 500K searches/s		X		X	X	X
Unstructured Data - 300+ PB e-mail/yr		X		X	X	X
Biometrics / Genetics - 1+ GB / person	X			X		X
Medical - 500+ TB/yr/hospital; 3GB/3D scan	X			X		X
Security - 1+PB surveillance video/city/day	X		X	X		X
Geospatial Data Mgmt - 350+ TB / company	X		X	X		X
Video Mgmt - 2.5+ B videos; 100+ M views / day	X	X	X	X	X	X

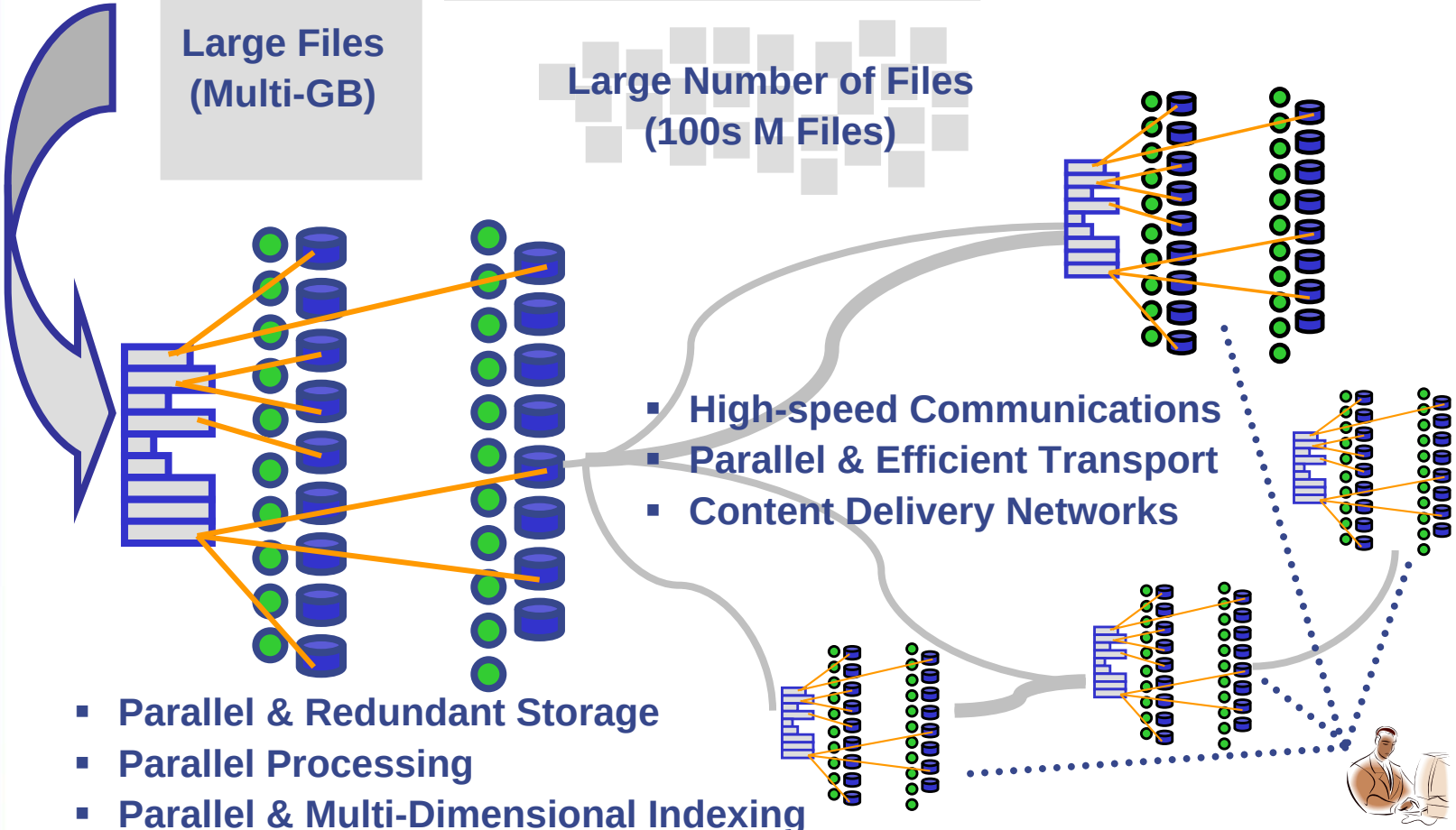
Increased Reliance on Clouds to Rapidly Refine Knowledge from Data Mountains

Emerging Cloud Data Management Capabilities

Large Files (Multi-GB)

Large Streams (Multi-GB)

Large Number of Files (100s M Files)



- Parallel & Redundant Storage
- Parallel Processing
- Parallel & Multi-Dimensional Indexing

- High-speed Communications
- Parallel & Efficient Transport
- Content Delivery Networks

- Ubiquitous Access

Toward User-Managed Information Factories[1]

Connecting People With Information

1 See WIRED, October 2006

Why Does it Matter?

Individual Scale

- **Scan 1 Terabyte on 1 disk:** **~ 3 hrs. @ 70 MB/s**
- **Send 1 Terabyte on user network:** **~ 90 days @ 1Mb/s + TCP**

Standard Enterprise Scale

- **Scan 1 Terabyte on 10 disks:** **~ 10 min. @ 125 MB/s**
- **Send 1 Terabyte on center network:** **~ 2 hrs. @ 1 Gb/s + TCP**

Cloud Scale

- **Scan 1 Terabyte on 1,000 disks:** **~ 6 sec. @ 125 MB/s**
- **Send 1 Terabyte on cloud networks:** **~ 1 min. @ 10 Gb/s + Efficient**

Typical Google “cell” leverages 3,000+ commodity nodes to support 500K accesses per second on 6+ PB of data.

A 1,000 commodity node Hadoop Cloud currently holds the Terabyte Sort Benchmark record.

Information Superiority is Time-Dependent

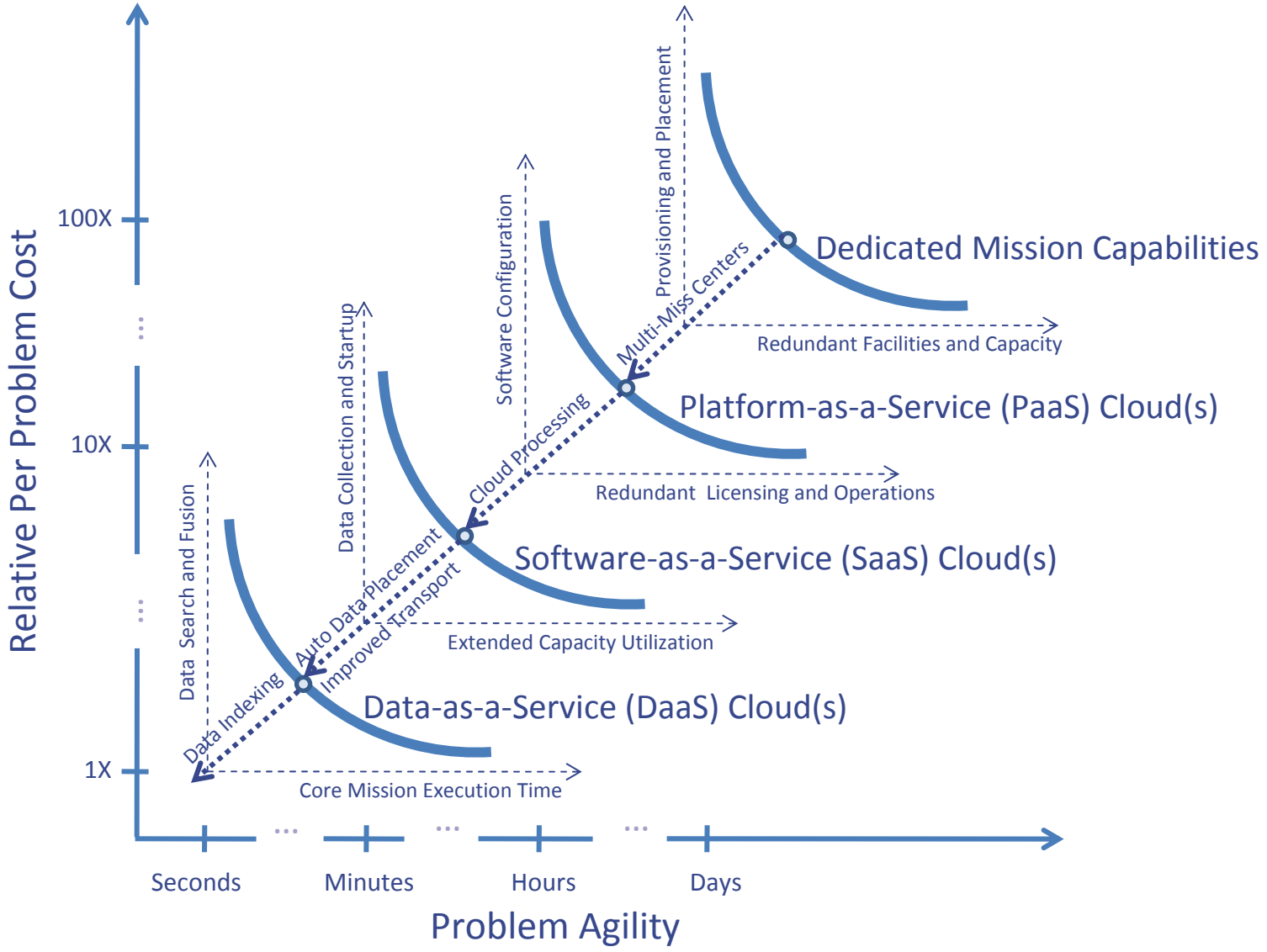
DoD Web-Scale Data Problems

Area	H-INT	G-INT	M-INT	S-INT	O-INT	Bio	Med	Log
Battlespace Awareness	N	S,F	N,F	N,S	N,S			
Force Application			N		N			
Force Protection						N	S,F	
Force Logistics								N

N = Large Number of Files
S = Large Data Streams
F = Large Files

Focus on Leveraging Commercial Synergies

DoD Clouds - What's the Business Case?



Cost-Effective Agile Response to Uncertainty

First Steps - DISA's RACE (Rapid Access Computing Environment) to PaaS & SaaS

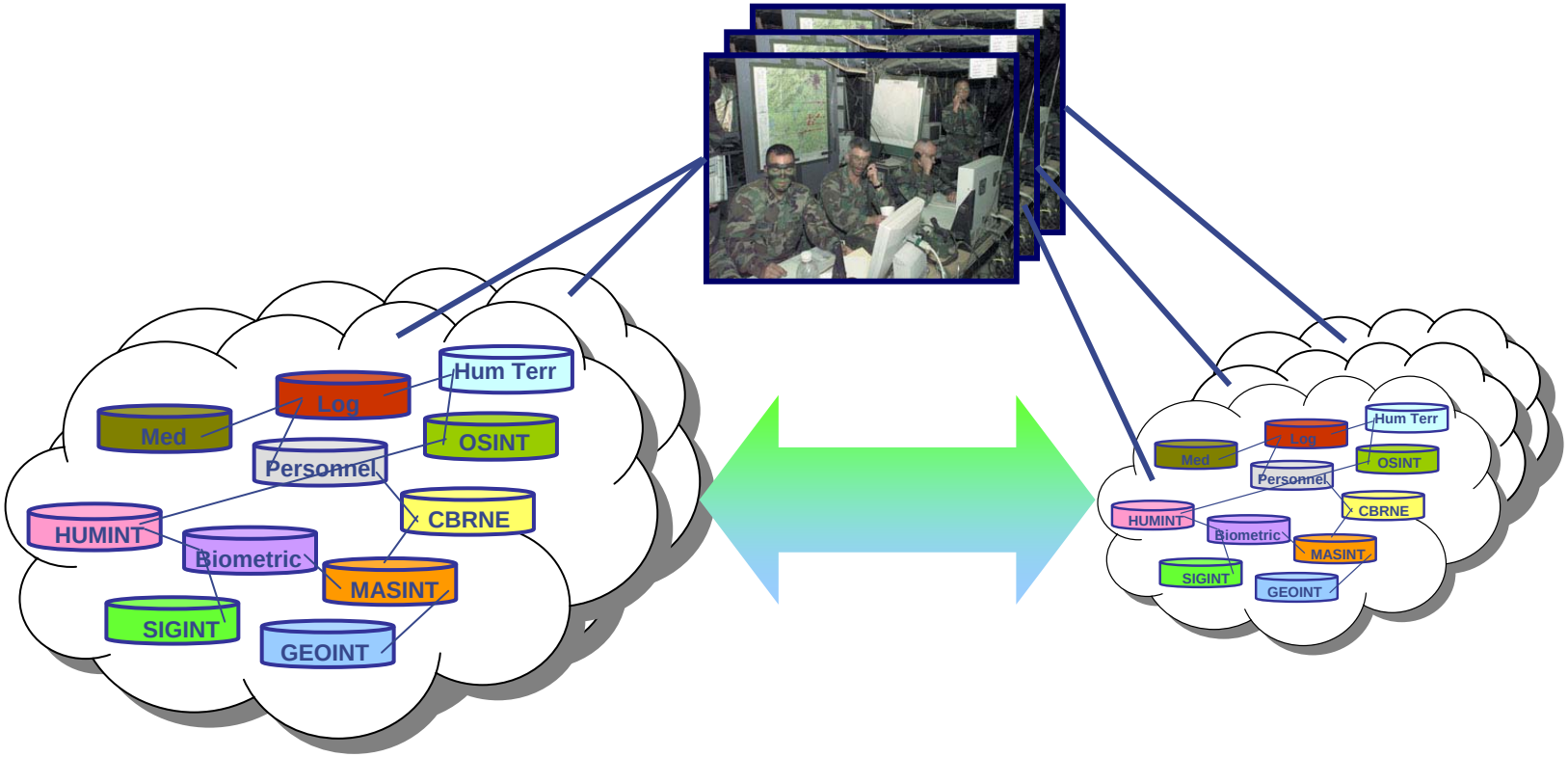
Objective	Approach	Results
• Rapid access to computing resources • Eliminate the need to procure physical infrastructure • Self service portal through a single, secure interface • User configurable server environments • Automated provisioning • Flexible billing options • Meet DoD security requirements	• HP C&I development of shared services utility for Rapid Access Computing Environment (RACE) • HP Server Automation and HP Operations Orchestration for provisioning and configuration management • HP Operations Manager for monitoring and control • HP Service Manager to automate incident & problem management • HP Systems Insight Manager and HP Proliant Essentials • Cluster Resources Moab for intelligent orchestration and Gold for billing	**Business outcomes** • Reduced costs • Consolidated simplified processes • Shortened time to delivery **IT improvements** • Flexible development platforms for Web, Application or Database • User can allocate own resources through Web interface • Can provision a server in a few minutes • CPU, memory, storage, virtual environment provided in one simple solution

DoD Dawn of the New Age of Petacomputing: Optimizing Processing, Storage, Bandwidth, Location, Power and Electricity[1]

1 See WIRED, October 2006

DaaS on the Horizon

It's about connecting the dots across the data … all the data, Responding to uncertainty in an agile fashion



"Develop a common "cloud" based on a single backbone network and clusters of servers in scalable, distributed centers where data is stored, processed and managed."[1]

1 See "DNI Vision 2015", July 2008