Zetta

1,000,000,000,000,000,000,000

The Design and Implementation of the Zetta Storage Service

October 27, 2009

Zetta's Mission

Simplify Enterprise Storage

Zetta delivers enterprise-grade storage as a service for IT professionals needing primary storage solutions

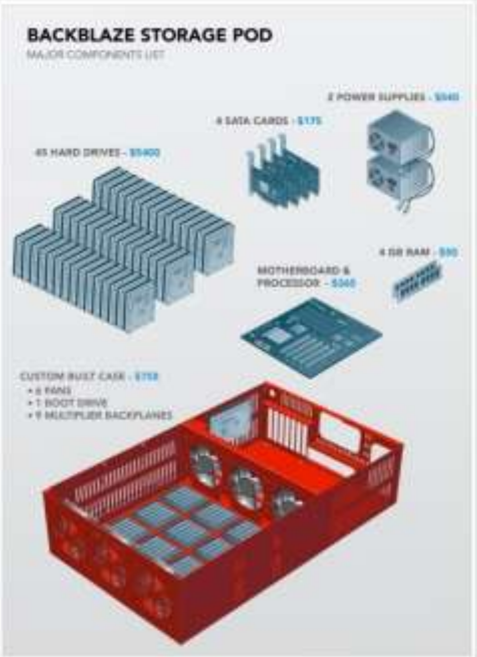
Market Landscape

- Tier 0/1 storage consumers (correctly) risk adverse– existing enterprises are not going to take mission critical transactional database and plug it into the cloud
- Network latency / speed of light an issue for many (not all) use cases
- Unstructured (file) data is majority of growth in terms of data footprint
- Lots of concerns about security, reliability, data integrity, etc, but examples of complete outsourcing of mission critical sensitive data (salesforce, email) also common
- CapEx, OpEx, and administration challenges colliding

Zetta Design Objectives

- Data Integrity
- Continuous Availability (failures, releases, scale out, moves, always consistent on disk)
- Strong consistency (respect sync()), be POSIX Compatible
- Multi Tenant (IO performance as well as footprint)
- Tiered Design, with independent horizontal scalability
- Economically Viable (commodity components)

A Tale of Two Boxes for “just” 40TB



Less internal redundancy, less hot swap, lower quality	Redundanty PSU, hot swap fans, higher quality
Internal Software Raid	Internal Software Raid
Less expensive purchase price	More expensive purchase price
False economy for most IT environments	Good choice for most IT environments

Beyond the Single Box

Approach	Positives	Negatives
Two “cheap” boxes, mirrored	more available & often less capital cost than one “expensive” box	What does the mirroring? Conflict resolution? Performance? Higher OpEx.
Application Partitioning	Great solution	Not applicable for many enterprise apps. Availability decreases as nodes increase (or cost increases for replication)
Distributed File Replication	Gets you past the “one box”	Complexity of the management layer, higher OpEx
Buy the Big SAN	Mature	Expensive, complex

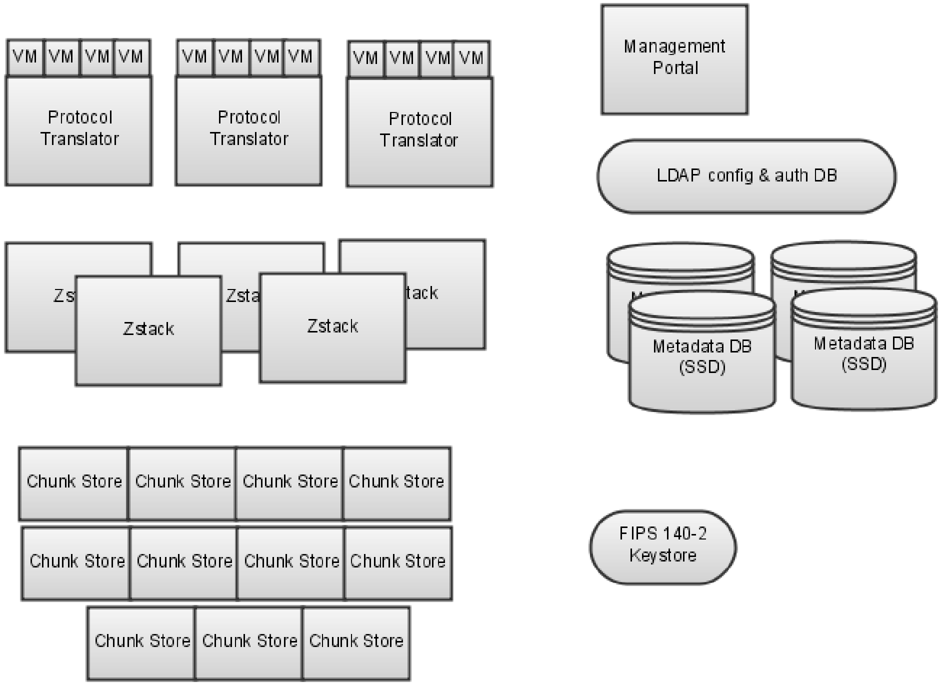
Distributed File Replication

- Hadoop
  - Data protection assumed to be at lower level (ie, raid card in every node) OR replicate every piece of data
  - Great for analytics, not a general purpose file system
- MogileFS
  - Open source, small install base, perl file-location-tracker
  - Replicates data for protection
  - No data integrity checking (hash/crc)
- Lustre
  - Data protection assumed to be at lower level (ie, raid card in every node)
- Parascale
  - Commercial – could we really run it “so much better,” that people would use us rather than run it themselves? (no.)
  - Data protection through replication
- **Overall**
  - **Software layer is complex, generally has single points of failure**
  - **Replication less space / opex / capital efficient than erasure coding**
  - **None designed for multi-tenancy**

Our Conclusion

In order to meet service objectives,
we can’t use off the shelf technology
and have to build something new.

The Implementation



ZettaFS Distributed File System

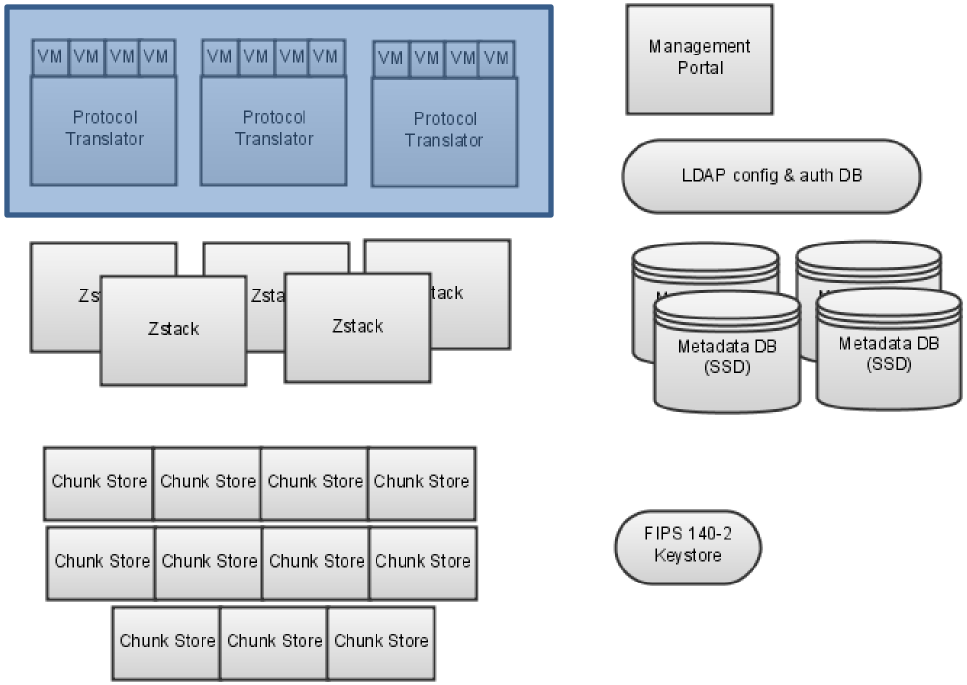
All elements implemented as network services

Centralized Metadata, holds ‘inode’ equivalents (on SSD)

10Gbps low latency ethernet

Basic unit of storage is a “chunk,” striped and protected across discrete nodes

The Implementation



Protocol Translator ==“NAS Head”

Xen VM-ZettaFS appears as local file system

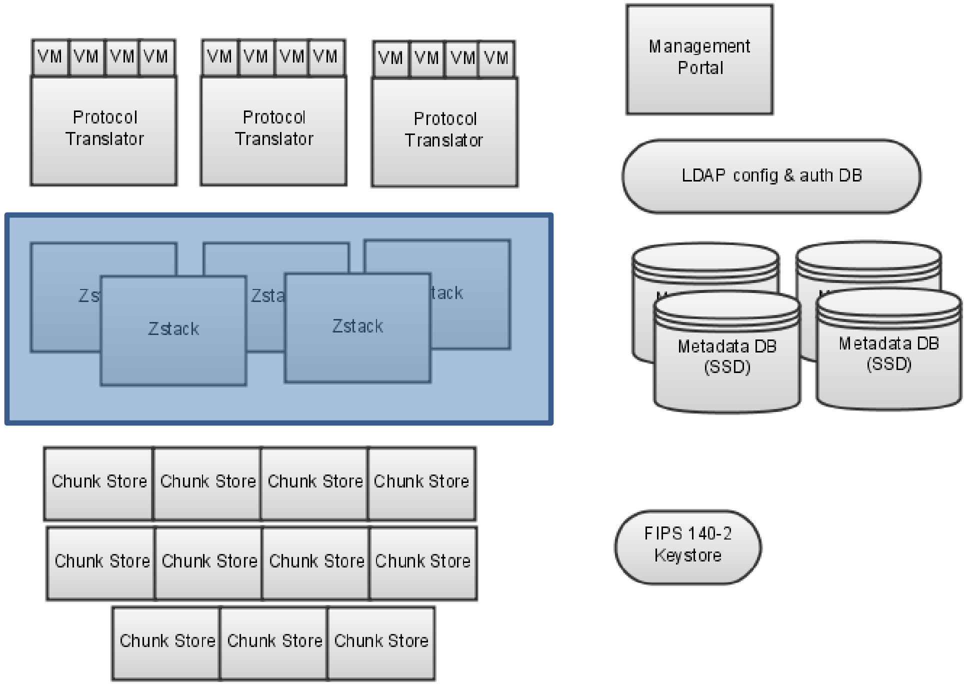
Pulls config and authentication creds from LDAP

QoS management

Caching

Reference Synchronization

The Implementation



Zstack ==“RAID Controller”

Reed-solomon chunk encoding / recovery

Write cache (local SSD & consensus quorum protocol)

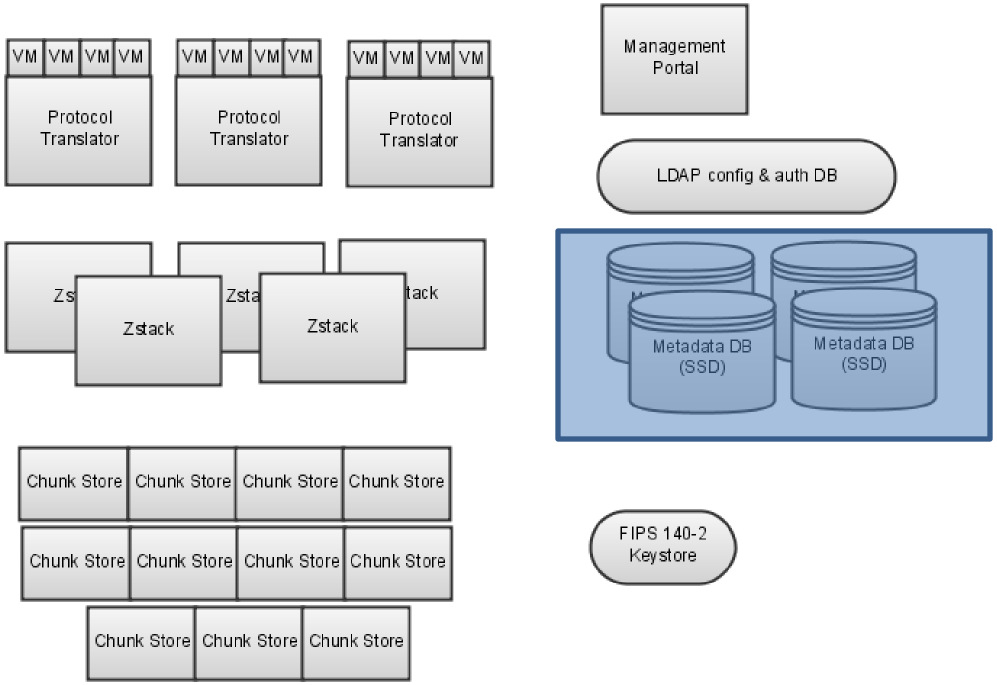
Metadata management

Lock Manager

Geo-Replication

Chunk placement rebalancing/optimization

The Implementation



Metadata DB

N+3 protection

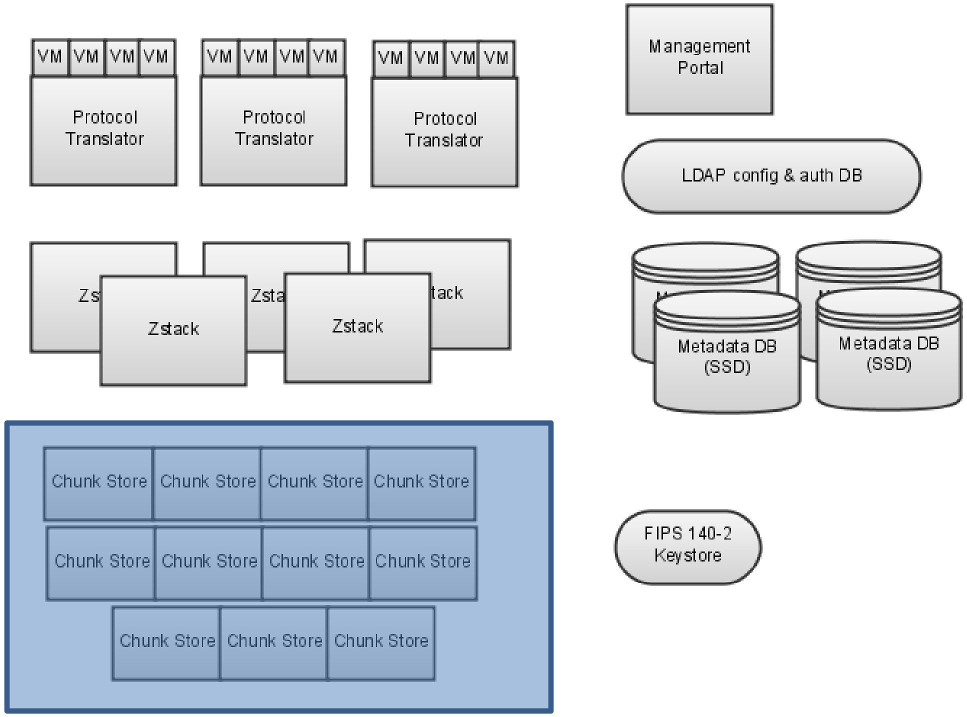
Volume -> file maps

File -> chunk maps

Raid stripe maps

Scalable / partitioned

The Implementation



Chunk Stores
== "Disks"

Caching Layer

Encryption / Decryption – 100% on-disk encryption

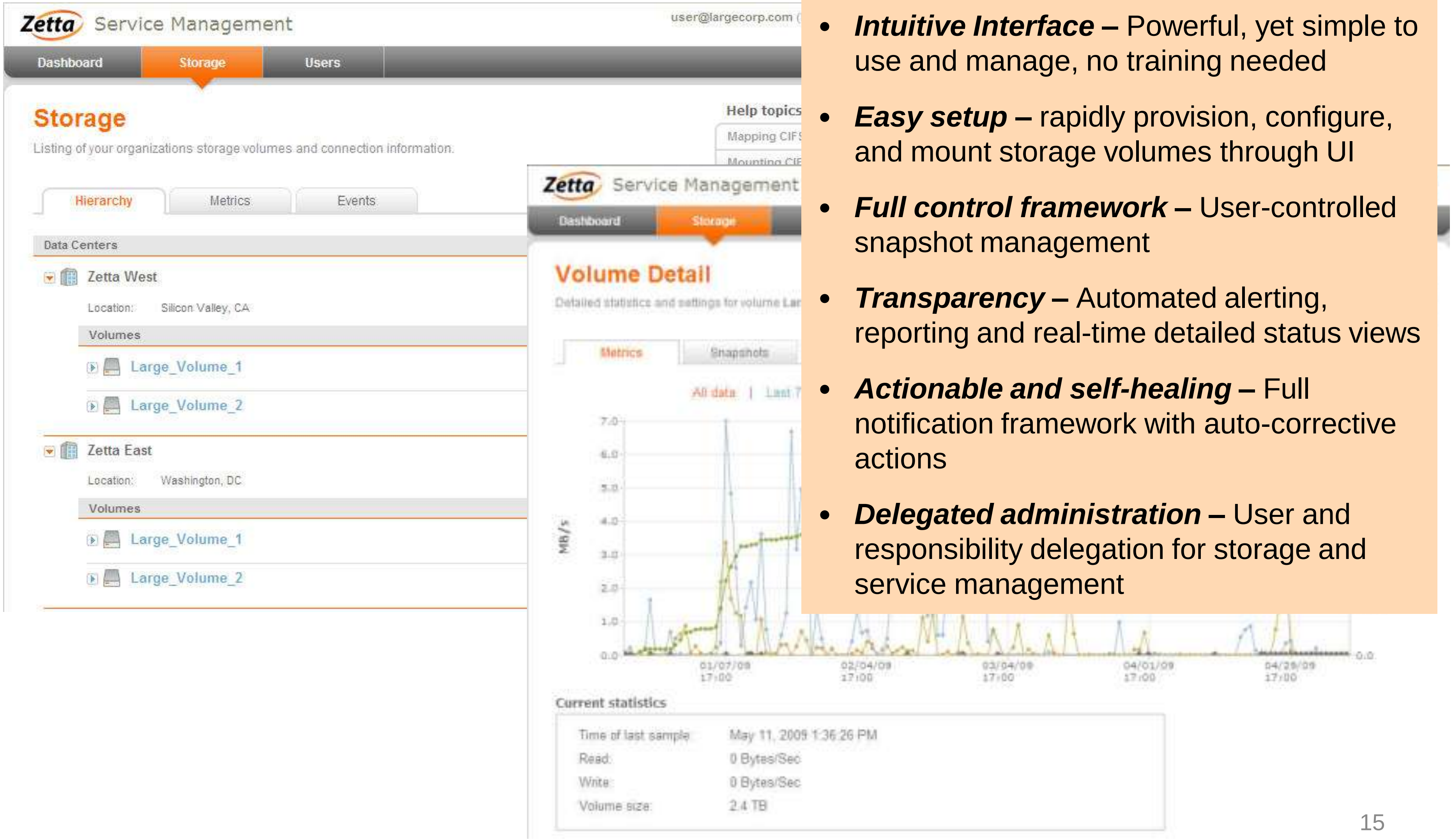
Background hash validation

Foreground read verification

Other Key Features

- Most NFS/CIFS requests are handled as metadata operations, which don’t require accessing the spindle layer
- Clustered mount capabilities
- Variable performance per volume – virtualize IO capacity, not just space
- Federated Authentication (LDAP)
- **Service Practices as important as technology**

System Management Portal



Storage

Listing of your organizations storage volumes and connection information.

Data Centers

Zetta West

Location: Silicon Valley, CA

Volumes

Large_Volume_1

Large_Volume_2

Zetta East

Location: Washington, DC

Volumes

Large_Volume_1

Large_Volume_2

Volume Detail

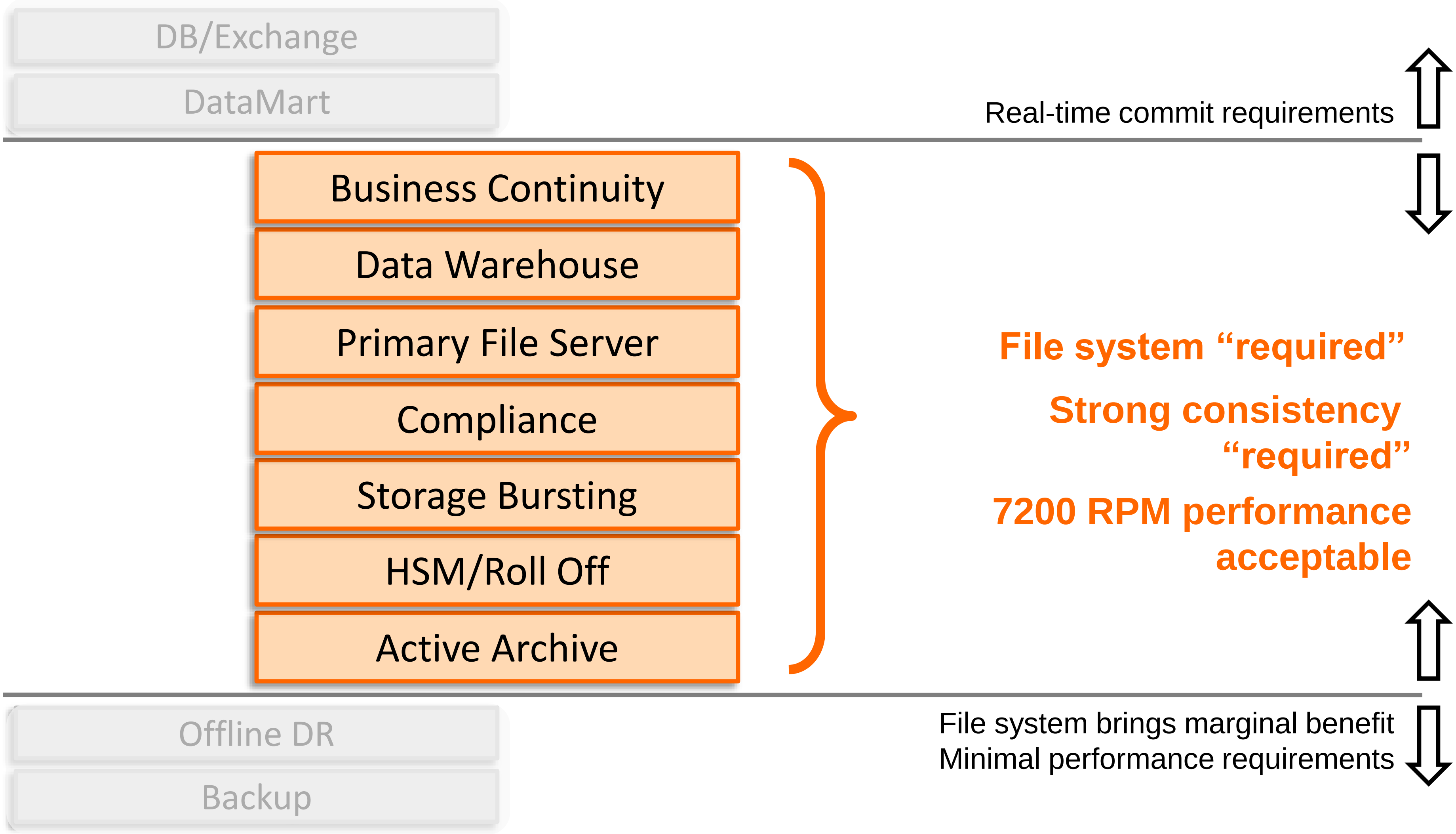
Detailed statistics and settings for volume Lar

Current statistics

Time of last sample:	May 11, 2009 1:36:26 PM
Read:	0 Bytes/Sec
Write:	0 Bytes/Sec
Volume size:	2.4 TB

- ***Intuitive Interface –*** Powerful, yet simple to use and manage, no training needed
- ***Easy setup –*** rapidly provision, configure, and mount storage volumes through UI
- ***Full control framework –*** User-controlled snapshot management
- ***Transparency –*** Automated alerting, reporting and real-time detailed status views
- ***Actionable and self-healing –*** Full notification framework with auto-corrective actions
- ***Delegated administration –*** User and responsibility delegation for storage and service management

Use Cases for Zetta Storage



DB/Exchange

DataMart

Real-time commit requirements ⇧

⇩

Business Continuity

Data Warehouse

Primary File Server

Compliance

Storage Bursting

HSM/Roll Off

Active Archive

File system "required"

Strong consistency "required"

7200 RPM performance acceptable

⇧

Offline DR

Backup

File system brings marginal benefit
Minimal performance requirements ⇩

Customer Proof Points

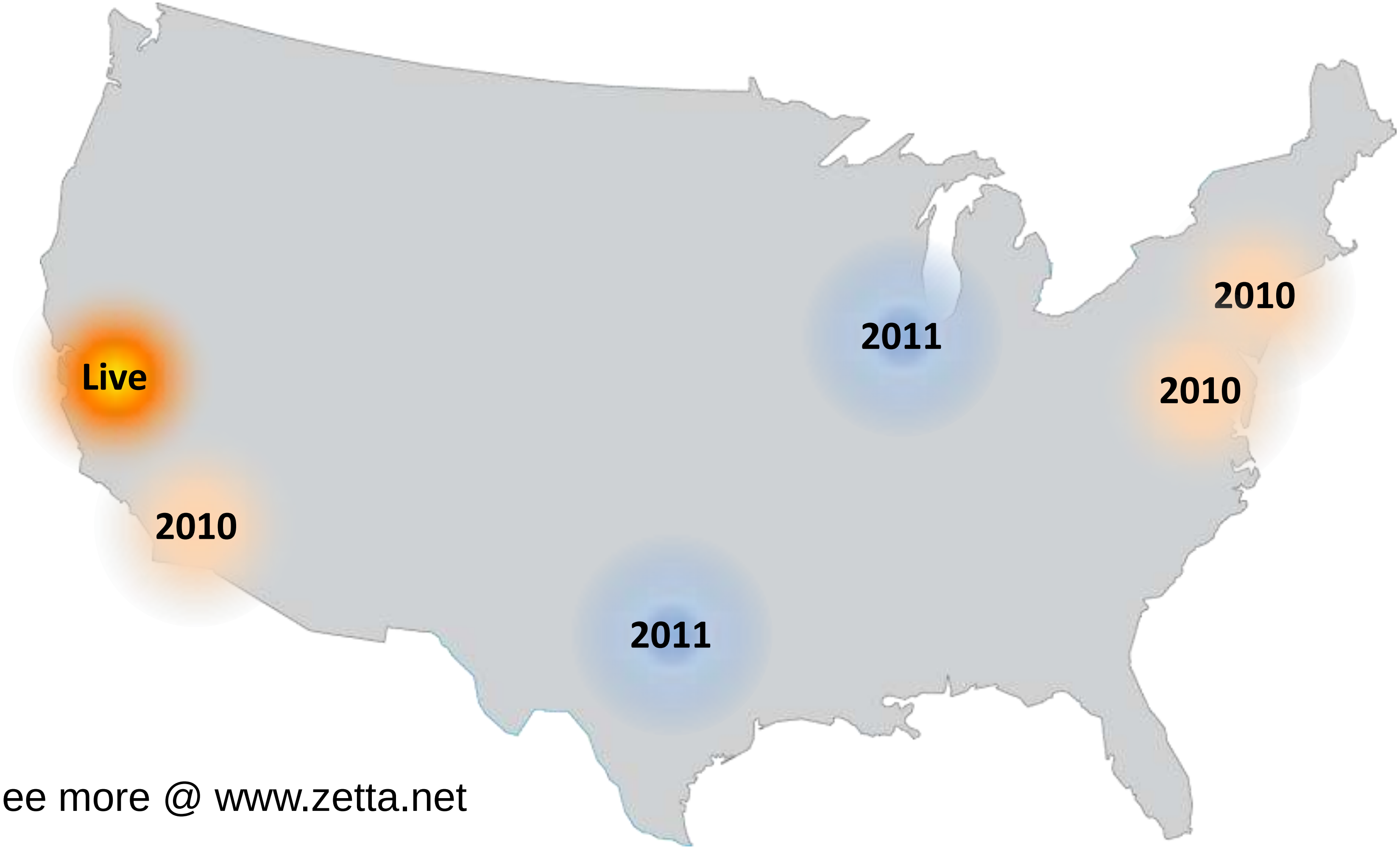
- Consumer/SMB IT/Media Services Provider
  - 4TB/day ingest
  - Expected Volume Size: 750T – 1 PB
  - Connected via 10Gbps Dedicated Circuit
- Large Silicon Valley Law Firm
  - Security, Data Integrity, SLA requirements
  - Need Snapshots, File System
  - Connected via Cross Connect
- Large Public University
  - Connected via Internet

Zetta Business Model

- Integrated Offering
  - All robust features included in base offering
  - Future protocols, APIs, performance improvement included
  - Customer service and support included
- One Simple Price
  - Starts @ $.25 per GB per month for 1 TB
  - Discounts for footprint volume, term
  - Connectivity options for lowest network % of TCO

Geo Expansion

Zetta



See more @ www.zetta.net