General Purpose Storage Servers

Open Storage Servers

Art Licht
Principal Engineer
Sun Microsystems, Inc
Art.Licht@sun.com
Agenda

- Industry issues and Economics
- Platforms
- Software
- Architectures
Industry Issues
aka
(Volume Economics)
“Hello, Bob? It’s your father again. I have another question about my new computer. Can I tape a movie from cable TV then fax it from my VCR to my CD-ROM then E-mail it to my brother’s cellular phone so he can make a copy on his neighbor’s camcorder?”
General Purpose Revolution

- It happened in compute over the last few decades
  - Volume CPUs killed special purpose hardware
- It's happening in storage right now
  - Custom hardware can't keep up with GP CPU
  - Custom OS's can't keep up with GP OS's (Solaris, Windows, Linux)
  - Custom interconnects can't keep up with Ethernet
- GP storage is the future
  - Low end is only the beginning
  - Every successful attack in IT is waged from below
- Market is ripe for disruption
General Purpose Advantages

- High throughput, high quality compute readily available
  - Lots of Threading: Multi-Core, Multi-Thread (MCMT)
  - Solaris built to manage MCMT (StarFire Legacy of late 90's).

- Up-leveled data management features
  - Enables cryptographic security, check sums for integrity, RAID

- Greater functionality in storage systems
  - Serving blocks transitions to “managing data and information”
    - Context Management, search, HSM, SLAs, objects, application level integration, etc..

- New breed of storage systems become possible
Server or Storage?

- Server
- Network and Storage Fabric
- Single Integrated System
- Multiple Disk Arrays
- Server + Storage Arrays
- Sun Fire X4500 Server
Volume Economics

Traditional Modular Storage

- 2 Single Core CPUs
- 16GB of Mem
- 48TB of Disk (SATA)
- ~$230k
- Does Not Include
  - Snapshot Licence
  - Replication Licence
  - Volume Copy License

Industry Standard Data Server

- 2 Quad Core Opteron CPUs
- 32GB of Memory
- 48TB of Disk (SATA)
- ~$62k
- Includes
  - Snapshot, Replication, Compression, Thin Provisioning
Volume drives pricing

Traditional Modular Controller

- 2 Single Core CPUs
- 8GB of Mem
- ~$50k

Industry Standard Server

- 2 Quad Core CPUs
- 48GB of Memory
- ~$12k
Storage Servers
Two ways to think about this

**Server Perspective**
Runs all my applications directly on the server closely coupled to huge disk store. Thumper is not covered by Storage Policy restrictions.

**Storage Perspective**
Compact Storage Array with the server included in the same enclosure. Very different to traditional storage devices.
Typical Applications

- Media Server
- Solaris 10 IB SRP
- Video Surveillance Capture
- Solaris 10 ZFS/NFS/CIFS Server/Storage
- VTL DeDup
- iSCSI Target Storage
- Bulk Storage
- Solaris 10 SAM-FS Archiving Server/Storage
- HPC Grid Lustre Server/Storage
- Mainframe Batch Processing
- Video Surveillance Capture
Data Warehouse Example

- **20 Terabytes**
  - 20 kW, 8 Racks
  - $20M

- **33 Terabytes**
  - 20 kW, 6 Racks
  - $7M

- **100 Terabytes**
  - 9 kW, 1 Rack
  - $1.8M

All pricing based on list.
The Value is in the Software

- Shared-nothing, massively parallel architecture
- Optimal for fast queries and loads
  - Shared-nothing, massively parallel architecture
  - Optimal for fast queries and loads
  - Each CPU core operates on its own distinct portion of the database
  - Every query uses all CPUs in the cluster
  - The DW100 delivers 1 TB per minute scan rates
  - **Mirroring/fault tolerance features**
Based on our experience with SAS SPDS the Sun Fire X4500 unit is an ideal platform for optimal SAS data storage - it provides parallel access to multiple drives while giving SPDS the ability to precisely control how those drives are used. Unlike larger (and much more expensive) storage solutions, the SunFire X4500 doesn't add multiple layers of management between SPDS and the spindles. This direct control allows the SPDS administrator to configure SPDS domains with a precision we haven't been seen before in a system with comparable capacity. **Our testing showed excellent performance, with read times in the range of 1GB/second.**

Combined with the unit's cost, the precise control and exemplary performance make the Sun Fire X4500 an SPDS "dream machine".

**SAS SPDS on Sun Fire X4500**
A Performance Study
Nick Welke, Zencos Consulting
SAS Quality Partner

Software
aka
Microcode
Open Solaris™ – Open, Built for Today, Architecture for Tomorrow

> Tremendous scale
  - Large physical memory and up to 256 cores/CPUs
  - ZFS - 128 bit file system architected for speed
  - SAM-Q provides a future-proof file format

> Dynamic management
  - Hot plug and predictive self healing
  - OS-level virtualization and Dtrace observability

> Extensive security
  - Evaluated to EAL 4+
  - Cryptographic framework and encrypted replication

> Guaranteed data integrity
  - Check summing and copy on write (recursive)
  - Double parity RAID-Z and hot spare devices
A New Way to Manage Data

End to End Data Integrity
Checksumming & Copy on Write
Easier Administration
Immense Data Capacity
World's 1st 128 Bit Filesystem

Features:
CDP, Replication, Encryption,
Thin Provisioning
The Worlds Best Storage Microcode
Hundreds of Systems
Thousands of ISVs and Partners results in “thick” Solaris
QFS/iSCSI Example

Data Area Network

Compute Nodes
NFS Clients

QFS Meta Data Servers

Data Servers

X4600

X4500 w/ZFS

NFS front end – QFS over iSCSI name space

Storage Developer Conference 2008
Copyright Sun Microsystems, Inc. Used
You Next Datacenter is Flat

- Traditional disk I/O networks are collapsed behind a networked Storage Server abstraction layer.
Very Large ZFS

Data Server Tree Architecture

A master ZFS Server aggregates SCSI volumes from a federation of lower ZFS storage into one large storage pool.

Individual ZFS Data Servers manages a page of drivers and expose these as iSCSI volumes.
ZFS

- ZFS core features
  - 128-bit capacity (256 quadrillion zettabytes)
  - Striping, Mirroring, RAID-Z, RAID-Z DP
  - Replication (1:N, N:1, manual, scheduled, continuous)
  - Snapshots (r/o and r/w, unlimited in number)
  - Compression (lowest CPU utilization, or 10 levels)
  - Thin provisioning for both filesystems and iSCSI LUNs

- Data Integrity
  - ZFS data, meta-data checksums (19 nines)
  - ZFS “ditto blocks” for important meta-data
  - Local mirroring, RAID-5, RAID-6, RAID-across-trays
Protocols

- CIFS
- NFS
- iSCSI
- FTP
- HTTP
- WebDAV
- FC
Storage Systems

Internal Storage

External Storage
HA
Shared Disk trays
Software Stack

Open Solaris

NexentaStor is a software based NAS and iSCSI solution with features that are superior to those of legacy hardware based NAS solutions, including unlimited incremental backups or 'snapshots', snapshot mirroring (replication), block level mirroring ('CDP'), integrated search, and the inherent virtualization, performance, thin provisioning and ease of use of the ZFS file system. Built upon our OpenSolaris distribution, NexentaStor adds several capabilities that are required by most organizations that result in improved reliability, manageability, and
General Purpose Storage Server

DataCenter Edition
- VM management + Gold support

Enterprise Edition
- software appliance + platform
  - NAS/SAN/iSCSI

Ubuntu
- #1 community + packaging

Solaris kernel
- multi-core + clustering

NexentaStor
- ease of use
- price / performance
- extensibility
- partner friendly

NexentaOS:
- Loves multiple cores
- Boot level ZFS
  - aka 'gnusolaris.org'

ZFS
- checksums
- not volumes
  - 128 bit

More on ZFS:
- universal: SAN/NAS/iSCSI
- performance: block size + prefetch
- Software RAID
I don't know the future. I didn't come here to tell you how this is going to end. I came here to tell you how it's going to begin.”

Thank You

Art Licht
Art.Licht@sun.com