ZFS
Data Management Optimization for Oracle Environments

Michael P. Sweeney
Oracle ACES Team Lead
Sun Microsystems, Inc.

March, 2007

Agenda

- What is ZFS?
- What's different about it?
- What can I do with it?
- How does it perform?
- What about Oracle ASM?
- Where does ZFS go from here?

What is ZFS?

A new way to manage data

End-to End Data Integrity

With check-summing and copy-on-write transactions

Immense Data Capacity

The world's first 128-bit file system

Easier Administration

A pooled storage model – no volume manager

Huge Performance Gains

Especially architected for speed

Trouble with Existing File Systems?

Good for the time they were designed, but...

No Defense Against Silent Data Corruption

Any defect in datapath can corrupt data... **undetected**

Difficult to Administer−Need a Volume Manager

Volumes, labels, partitions, provisioning and lots of limits

Older/Slower Data Management Techniques

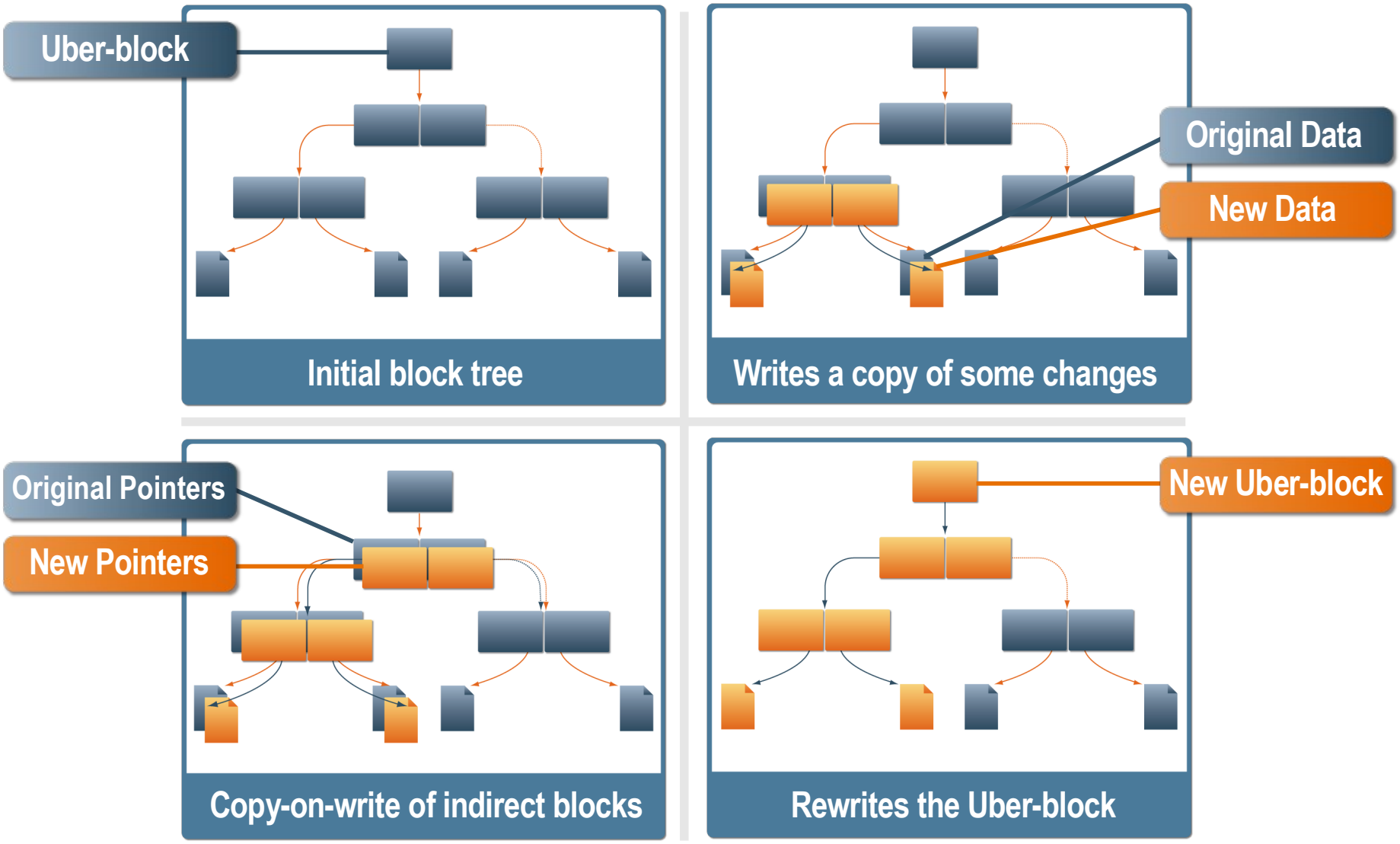
Fat locks, fixed block size, naive pre-fetch, dirty region logging

DATA
INTEGRITY

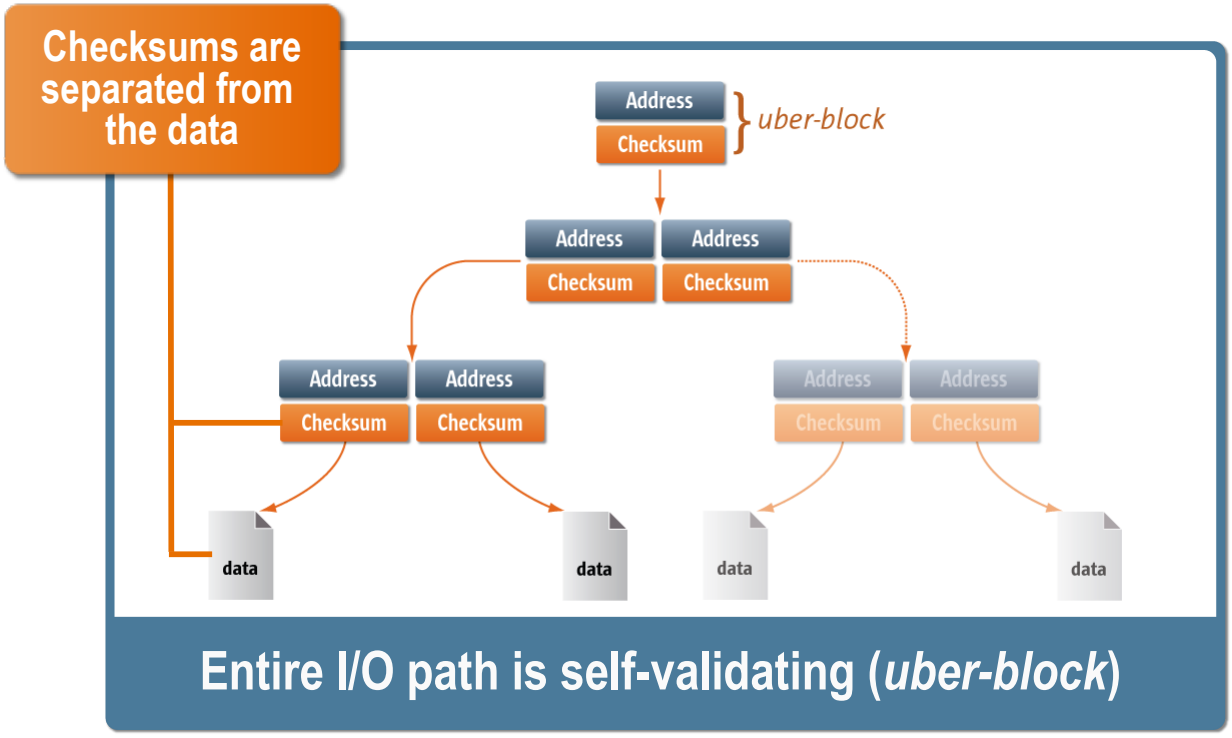
ZFS Data Integrity Model

- Copy-on-write, transactional design
- Everything is checksummed
- RAID-Z/Mirroring protection
- Disk Scrubbing

Copy-on-Write and Transactional



End-to-End Checksums



Prevents:

- Silent data corruption
- Panics from corrupted metadata
- Phantom writes
- Misdirected reads and writes
- DMA parity errors
- Errors from driver bugs
- Accidental overwrites

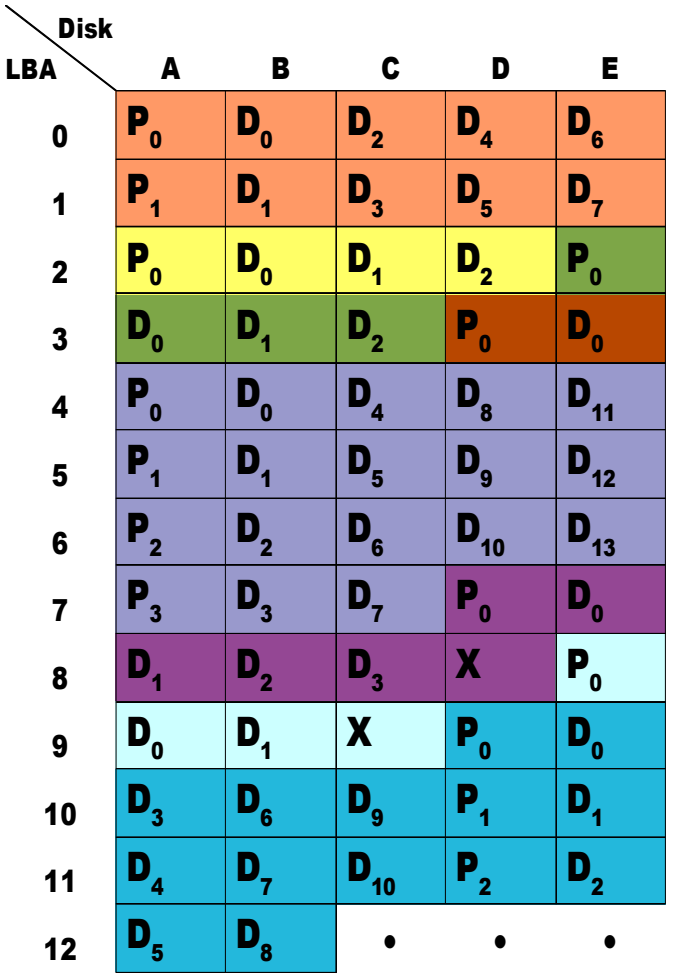
RAID-Z Protection

RAID-5/6 and More

- ZFS provides better than RAID-5 availability
- Striping uses dynamic widths
  - Each logical block is its own stripe
- All writes are full-stripe writes
  - Eliminates read-modify-write (So it's fast!)
- Eliminates RAID-5 “write hole”
  - No need for NVRAM
- Need more protection?
  - RAID-6 (double parity)

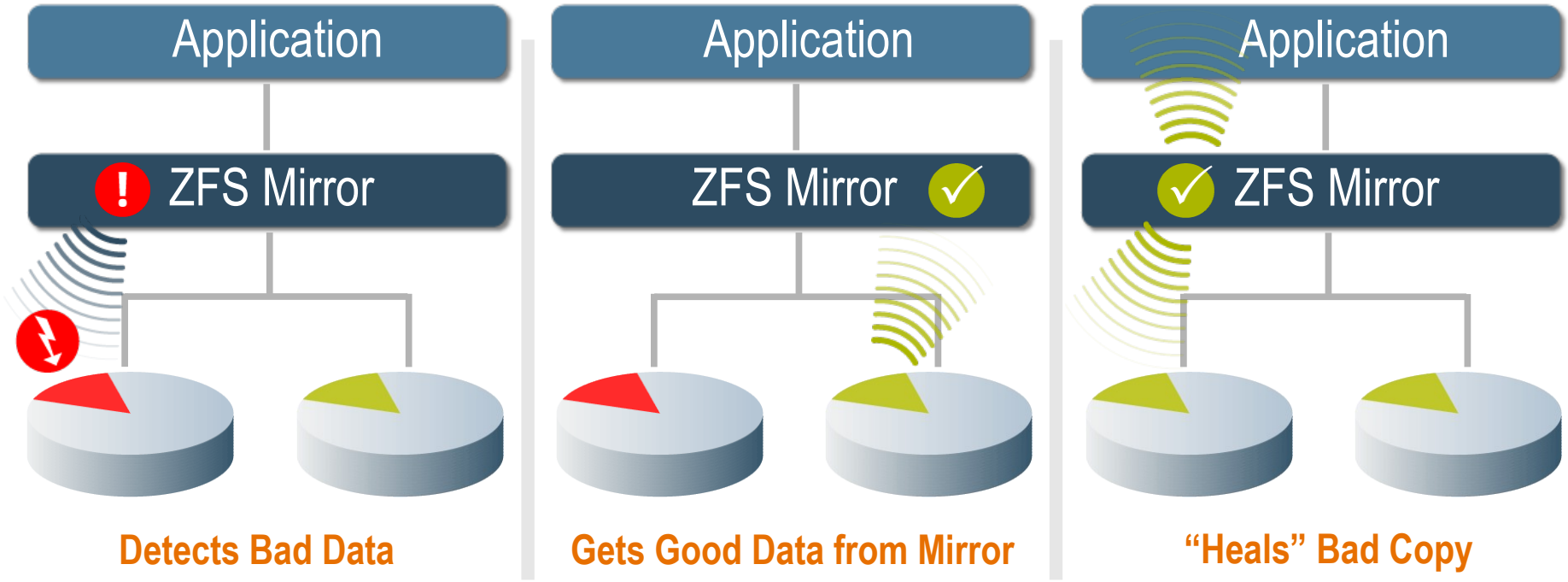
RAID-Z

- Dynamic stripe width
  - > Variable block size: 512 – 128K
  - > Each logical block is its own stripe
- Both single- and double-parity
- All writes are full-stripe writes
  - > Eliminates read-modify-write (it's fast)
  - > Eliminates the RAID-5 write hole (no need for NVRAM)
- Detects and corrects silent data corruption
  - > Checksum-driven combinatorial reconstruction



Self-Healing Data

ZFS can detect bad data using checksums and “heal” the data using its mirrored copy.



Disk Scrubbing

- Uses checksums to verify the integrity of all the data
- Traverses metadata to read every copy of every block
- Finds latent errors while they're still correctable
- It's like ECC memory scrubbing – but for disks
- Provides fast and reliable re-silvering of mirrors

128-bit File System

No Practical Limitations on File Size, Directory Entries, etc.

Concurrent Everything

Immense Data Capacity

EASIER
ADMINISTRATION

Easier Administration

- Pooled Storage Design makes for Easier Administration

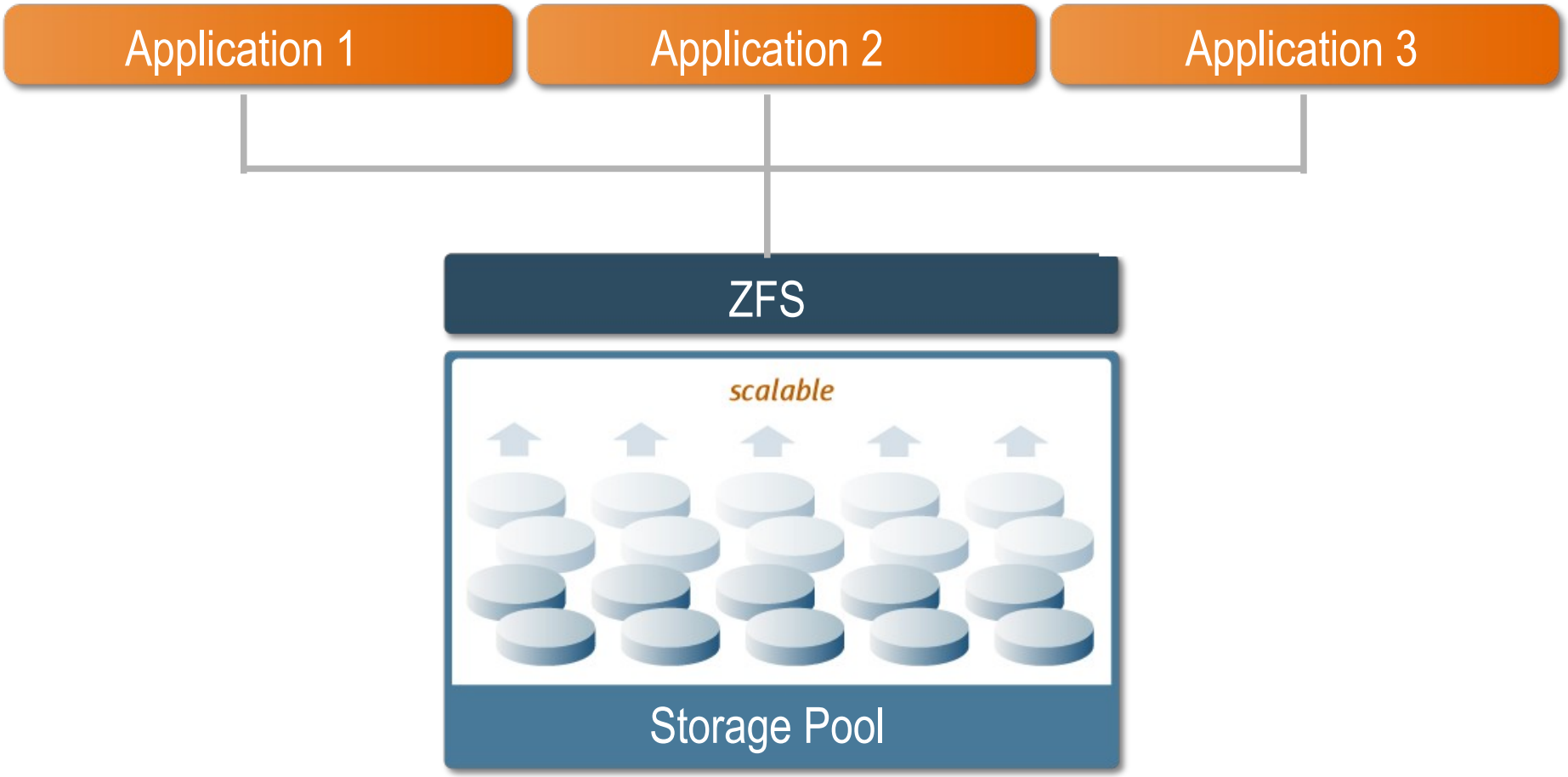
No need for a Volume Manager!

- Straightforward Commands and a GUI
  - > Snapshots & Clones
  - > Quotas & Reservations
  - > Compression
  - > Pool Migration
  - > ACLs for Security

No More Volume Manager!

Automatically add capacity to shared storage pool

"zpool add tank mirror c1t0d0 c1t1d0"

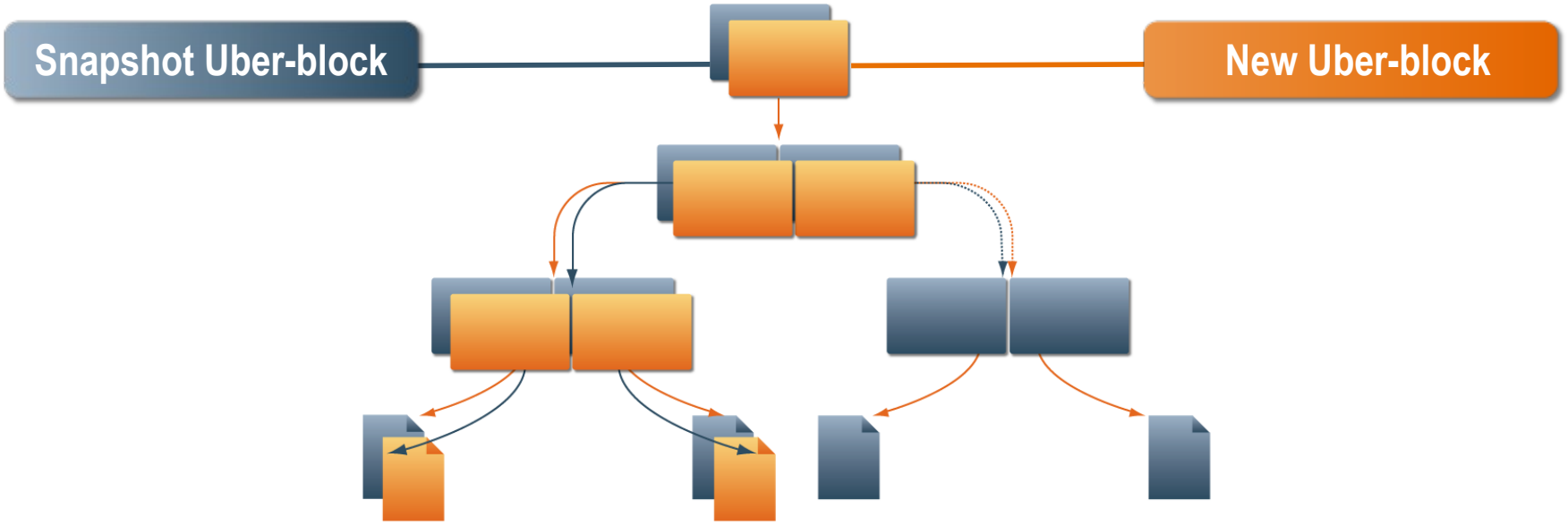


ZFS File systems are Hierarchical

- File system properties are inherited
- Inheritance makes administration a snap
- File systems become control points
- Manage logically related file systems as a group

ZFS Snapshots

- Provide a read-only point-in-time copy of file system
- Copy-on-write makes them essentially “free”
- Very space efficient – only changes are tracked
- And instantaneous – just doesn't delete the copy



Atomic ZFS Snapshots

Snapshot an entire file system hierarchy

Ensures all file systems are consistent at time of snapshot (all or nothing approach)

Example: Snapshot all file systems starting at tank/home

```
# zfs snapshot -r tank/home
```

ZFS Clones

Writable copy of a snapshot

Ideal for storing many private copies of shared data:

- Software installations
- Workspaces
- Diskless clients

Example: Create a clone of your OpenSolaris source code

```
# zfs clone tank/solaris@monday tank/ws/lori/fix
```

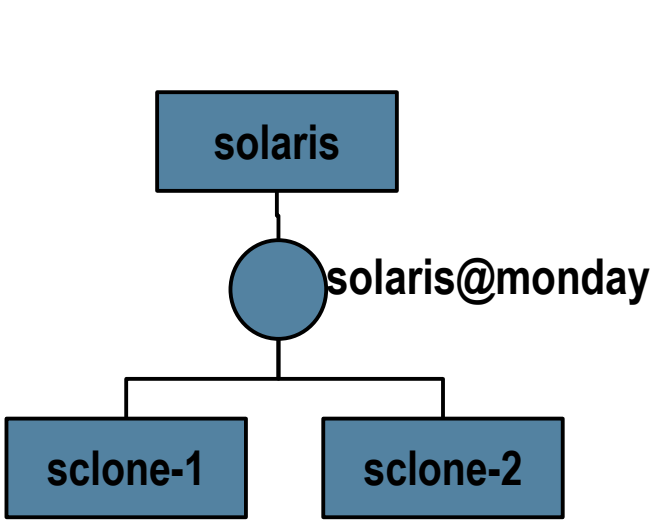
ZFS Clone Promotion

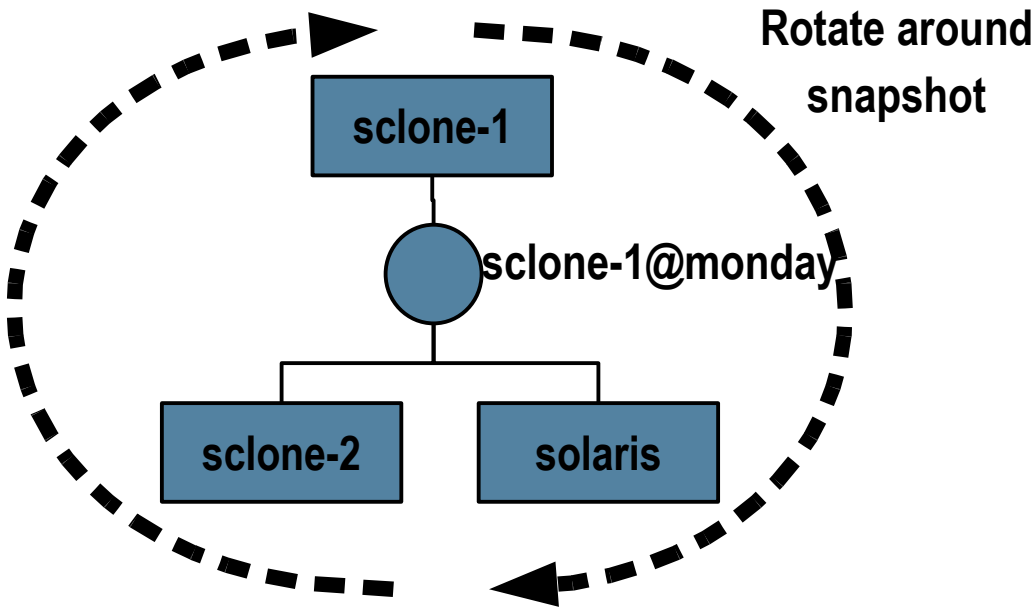
Changes the relationship between parent and child

Replaces a ZFS file system with a clone

Example: Replace source code with an updated clone

```
# zfs promote tank/sclone-1
```





Quotas and Reservations

- To control pooled storage usage, administrators can set a quota on a per file system basis

> Limit Tim to a quota of 10g

```
# zfs set quota=10g tank/home/tim
```

- Or they can set a *reservation* (minimum)

> Guarantee Fred a reservation of 20g

```
# zfs set reservation=20g tank/home/fred
```

“Adaptive Endian-ness”

- Hosts always write in their native “endian-ness”

Opposite “Endian” Systems

- Write and copy operations will eventually byte swap all data!

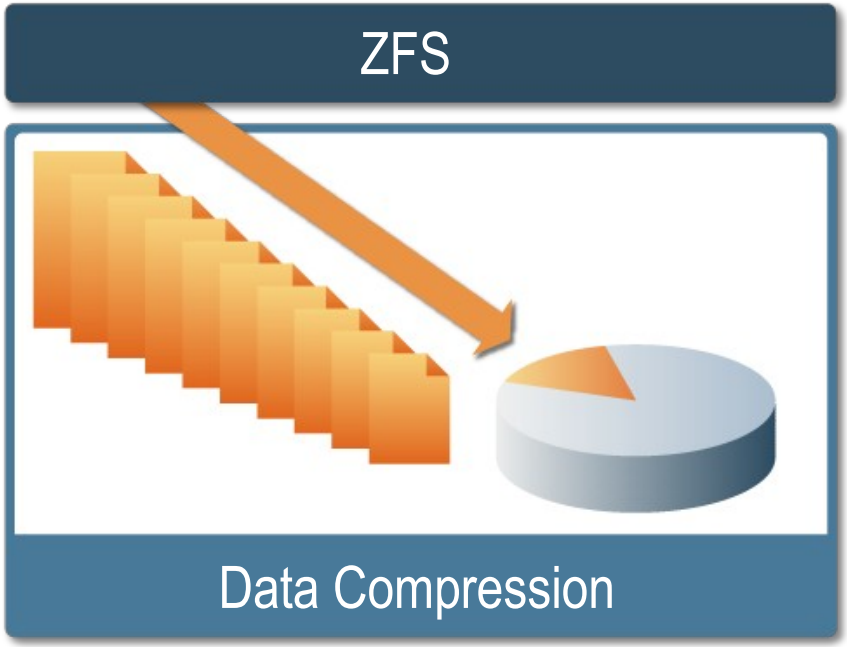
Config Data is Stored within the Data

- When the data moves, so does its config info

Storage Pool Migration

Data Compression

- Reduces the amount of disk space used
- Reduces the amount of data transferred to disk – increasing data throughput

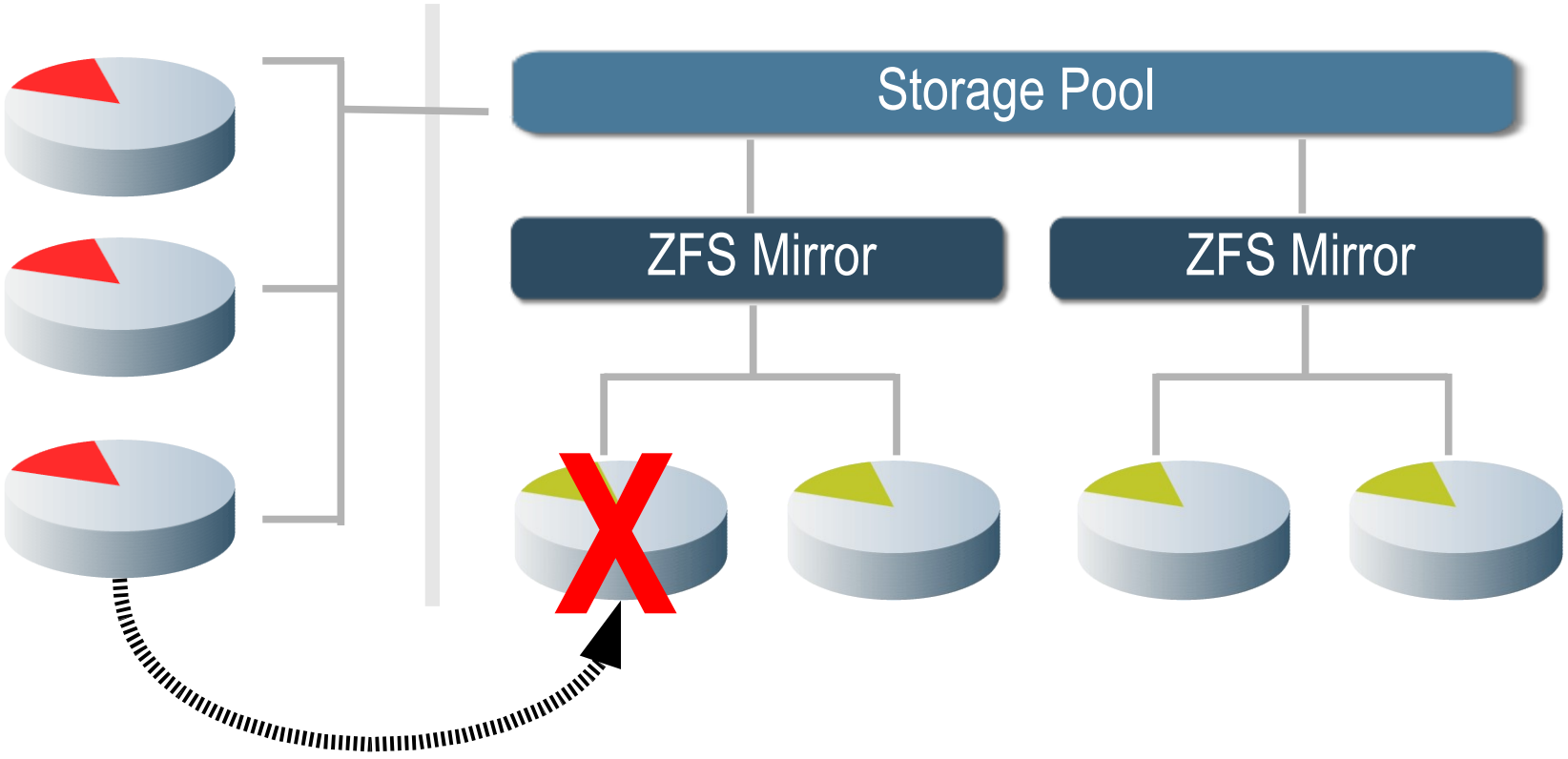


Reliability

FEATURES

Hot Spares

Automatically replaces faulted drive with hot spare device



ZFS Integrated with Fault Manager

- Errors are reported to FMA
- Checksum, I/O, device and pool errors are logged
- Integrated with Hot Spare Functionality

Ditto Blocks

- Data replication above and beyond mirror/RAID-Z
  - > Each logical block can have up to three physical blocks
    - Different devices whenever possible
    - Different places on the same device otherwise (e.g. laptop drive)
  - > All ZFS metadata 2+ copies
  - > Coming soon: settable on a per-file basis for precious user data
- Detects and corrects silent data corruption
  - > If the first copy is missing or damaged, try the ditto blocks
  - > In a multi-disk pool, ZFS survives any non-consecutive disk failures
  - > In a single-disk pool, ZFS survives loss of up to 1/8 of the platter
- ZFS survives failures that send other filesystems to tape

BREATHTAKING

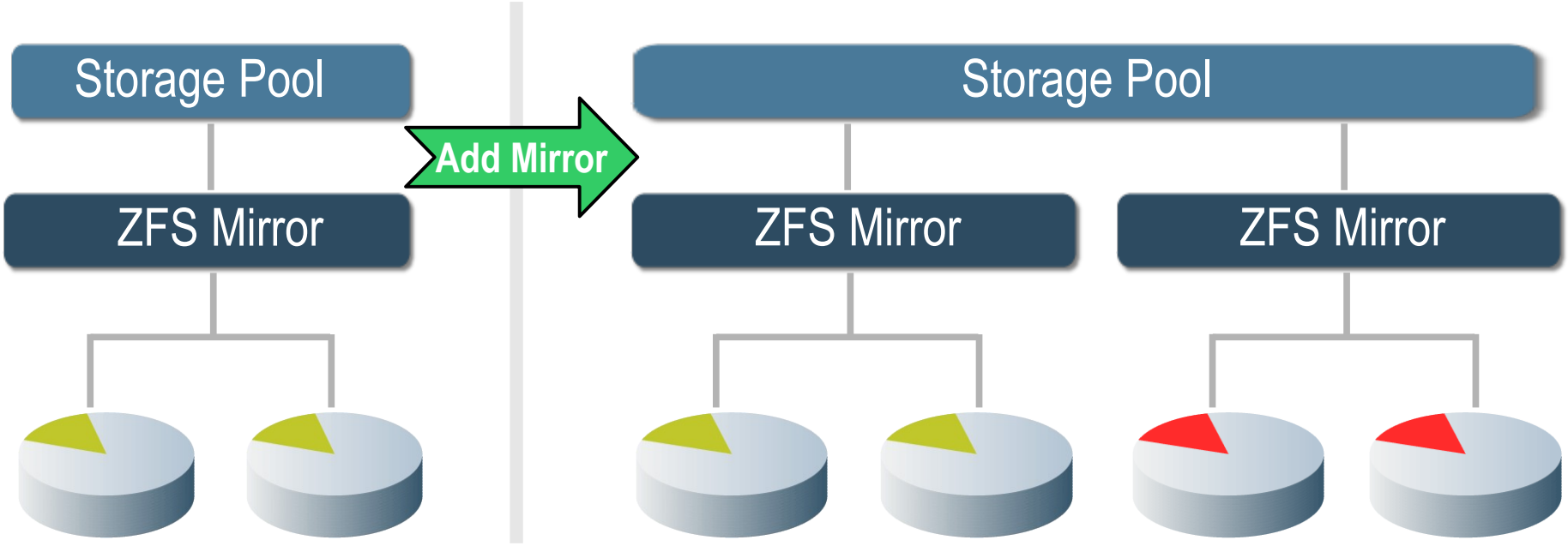
PERFORMANCE

- Copy-on-Write Design
- Multiple Block Sizes
- Pipelined I/O
- Dynamic Striping
- Intelligent Pre-Fetch

Architected for Speed

Dynamic Striping

ZFS can add bandwidth and capacity to all Filesystems



What About Oracle ASM?

- ZFS is not a Clustered Filesystem – YET
- Oracle ASM is limited to Raw Devices Only
- For maximum availability, use SunCluster Advanced Edition for Oracle 10gRAC
  - >This solution contains a QFS license
  - >QFS is cluster-aware - multi-reader/writer

Generic Configuration Recommendations for Oracle

- Separate zpools for Data and Redo Logs
- Match zfs block size to Oracle db_block_size
  - Data – 8k works well
  - Redo – 128k
- For small random reads, mirroring is best (disks are cheap)

Generic Configuration Recommendations for Oracle

- ZIL (zfs intent log) tuning for Tier 1 Storage
  - Will help write performance on high-end storage with lots of cache
  - fsflush algorithm optimized for cheap JBOD
  - Use whole disks rather than slices where possible

Check the Sun Blogs -

- ZFS Performance Blog
  - http://blogs.sun.com/roch/entry/tuning_the_knobs
- ZFS OLTP Workloads
  - http://blogs.sun.com/realneel/entry/zfs_and_databases_time_for
- General Microbenchmarks (spec.org, OASB)
  - http://blogs.sun.com/mrbenchmark

ZFS Command Summary

- zpool
  - (create, destroy, add, remove, iostat, import, export, scrub, upgrade)

http://docs.sun.com/app/docs/doc/819-2240/6n4htdnou?a=view

- zfs
  - (create, destroy, mount, send, receive, snapshot, rollback, set, get, list, share)

http://docs.sun.com/app/docs/doc/819-2240/6n4htdnop?a=view

ZFS Operating System Support

- Solaris 10 Update 3 (11/06)
  - On SPARC
  - On X64 (AMD and Intel)
- FUSE/Linux
  - Community Port w/Limited Feature Support
- Debian? RedHat? - Stay Tuned
  - Ian Murdock Works for Sun (LSB – Debian Founder)

Cost and Source Code

ZFS is FREE*

*Free
$ USD0
€ EUR0
£ GBP0
kr SEK0
¥ YEN0
元 YUAN0

opensolaris

- ZFS source code is included in Open Solaris
  - > 47 ZFS patents added to CDDL patent commons

And for the Future

More Flexible

- Pool resize and device removal
- Booting / root file system
- Delegated Users
- NFS Client snapshots
- iSCSI Integration

More Secure

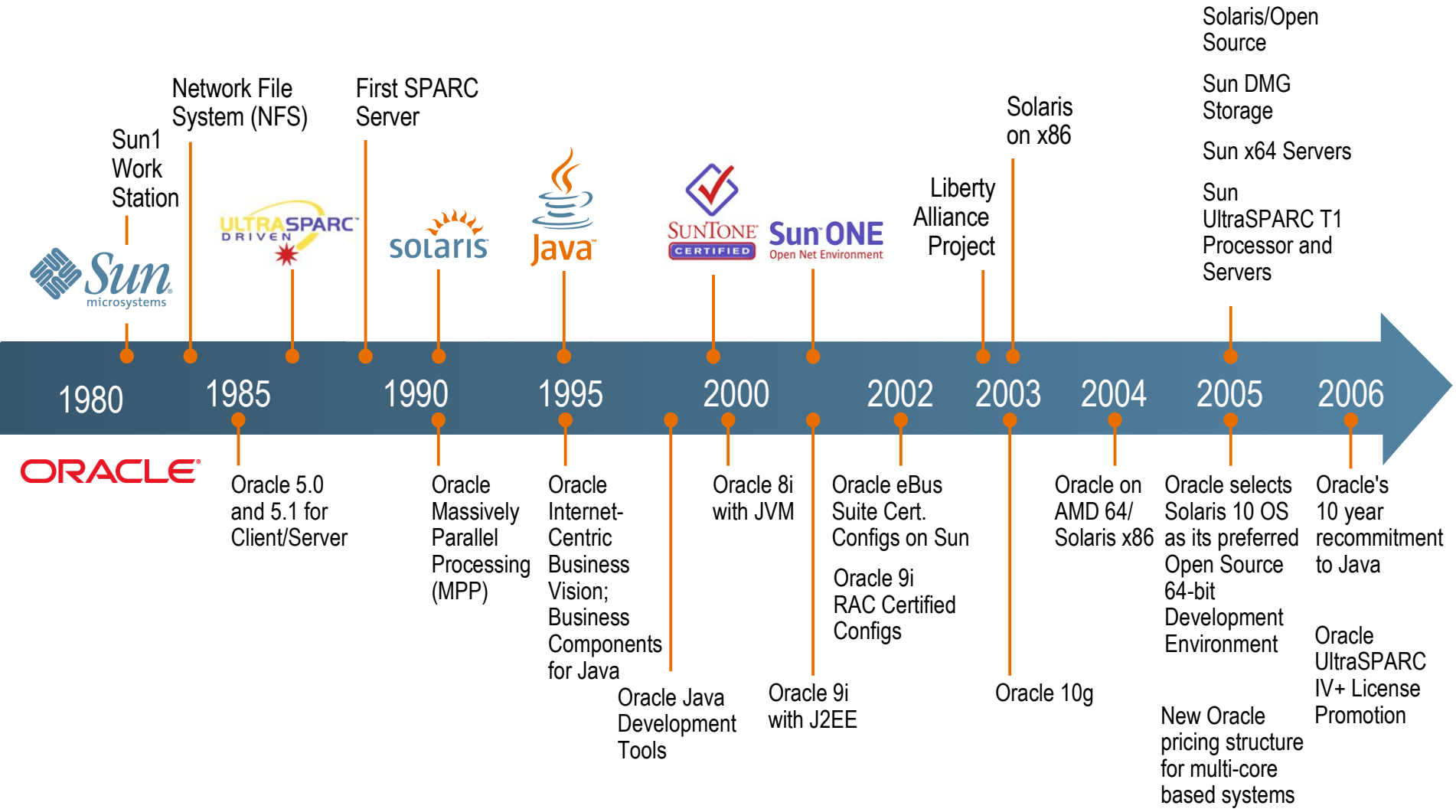
- Encryption
- Secure delete — overwriting for “absolute” deletion

More Reliable

- DTrace providers
- Track configuration changes ('zpool history')
- Improved write failure handling
- Corruption identification

Sun and Oracle Roadmap:

Two Decades of Partnership Alignment



Sun's Oracle Resources

Global Resources Working with Oracle

- Sun Solution Center for Oracle
- Market Development and Engineering
- Oracle ACES Field Architects (Me)
- Performance Application Engineering and OEM
- Business Applications Consulting
- VOS (Veritas-Oracle-Sun) and Sun VIP
- Systems Practice Dedicated Resources
- Storage Practice Dedicated Resources

Of course, Oracle has dedicated Sun resources...

Sun and Oracle Together

ORACLE®

Sun microsystems

- 10-year extension of our Java agreement
- Oracle chooses Solaris 10 as its preferred 64-bit development and deployment platform
- New Oracle licensing price advantages for multicore processors
- 85% of Sun's Customer's Run Oracle, and Vice-Versa
- Sun is the #1 market share leader with Oracle.

Sun Runs on Oracle, Oracle Runs on Sun

- Sun IT – Sun RAMP global consolidation project:
  - Consolidated 100 servers to one
  - 60 legacy apps in 55 countries to one single Oracle global instance
  - Reduces costs, simplified global processes, scalable application infrastructure
  - Only major hardware provider running its business on Oracle

- Oracle IT:
  - Sun is the primary IT system for single global instance in 150 countries
  - 70 ERP databases consolidated to 1 ERP database
  - Database backend, 4 x SunFire F25k's, running RAC and Sun Cluster

Oracle's Global Single Instance (GSI) Grid

Austin Data Center
Texas, USA

4 x

E25K

+

solaris

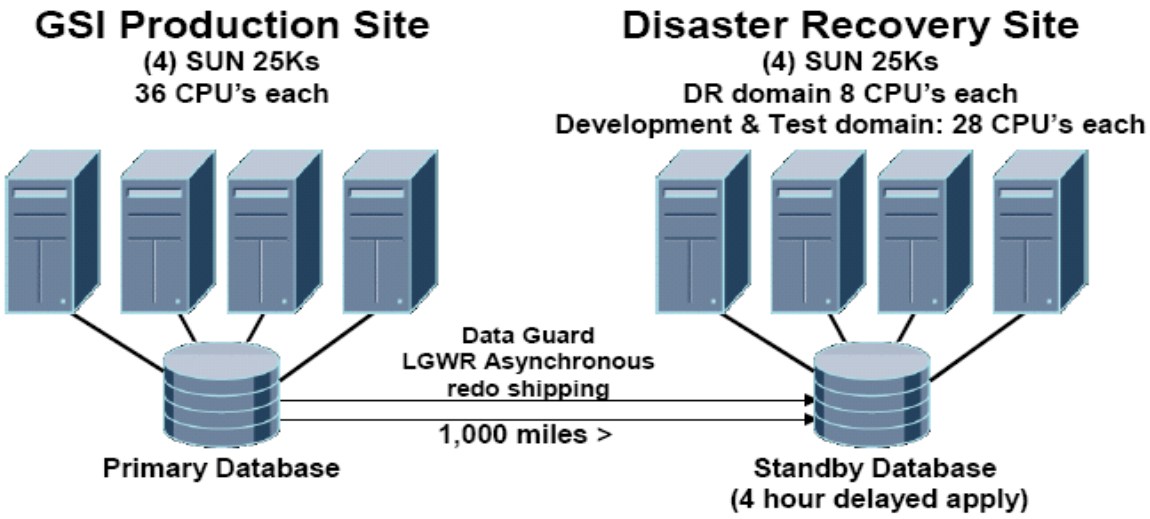
Rocky Mountain
Data Center
Colorado Springs,
Colorado, USA

4 x

E25K

Oracle Corp.
Global Single Instance

- 24*7 Business Critical
  - Accessed worldwide
- Peak loading
  - 7000 concurrent users
  - 225 concurrent batch jobs
- 7 Tb data
- SLA
  - 15min recovery time
  - No DR overhead on primary server
- Automated management



ZFS
Data Management Optimization for Oracle Environments

Michael P. Sweeney
Oracle ACES Team Lead
Sun Microsystems, Inc.

mike.sweeney@sun.com
(770) 360-6436