e&m computing Ltd.

א.מ.ת מיחשוב בע"מ

ZFS - The File System of the Future

Yitschak Goldberg - InterServe

Agenda

- What is ZFS?
- Why a new file system?
- What's different about it?
- What can I do with it?
- How much does it cost?
- 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

Easier Administration

A pooled storage model – no volume manager

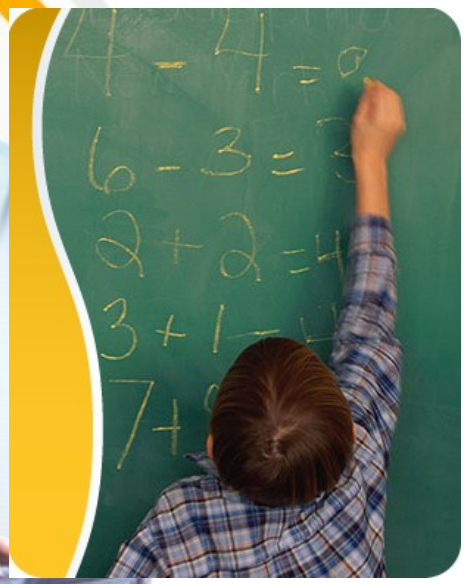
Immense Data Capacity

The world's first 128-bit file system

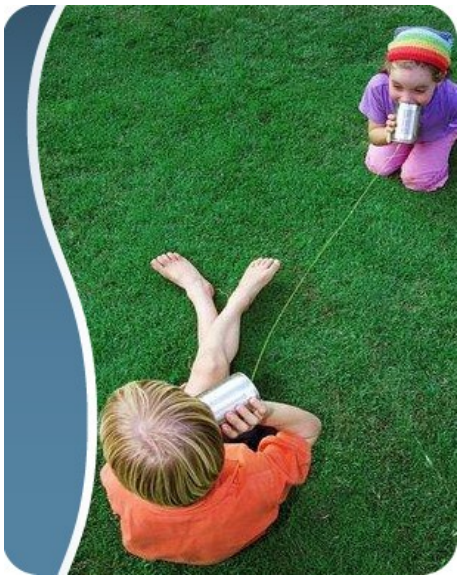
Huge Performance Gains

Especially architected for speed

Why a New File System?



Data Management Costs are High



The Value of Data is Becoming Even More Critical

The Amount of Storage is Ever-Increasing

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 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

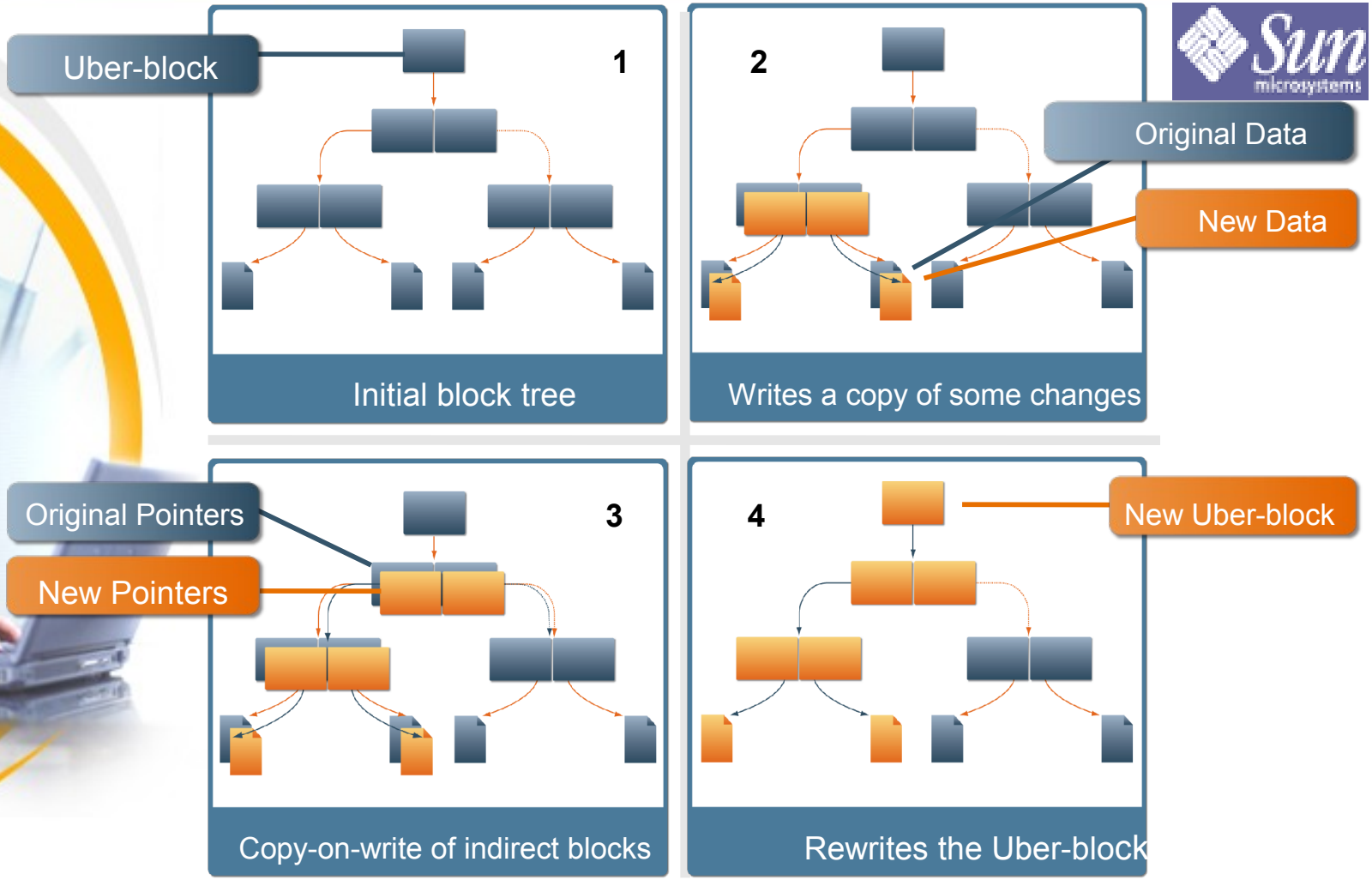
- DATA

INTEGRITY

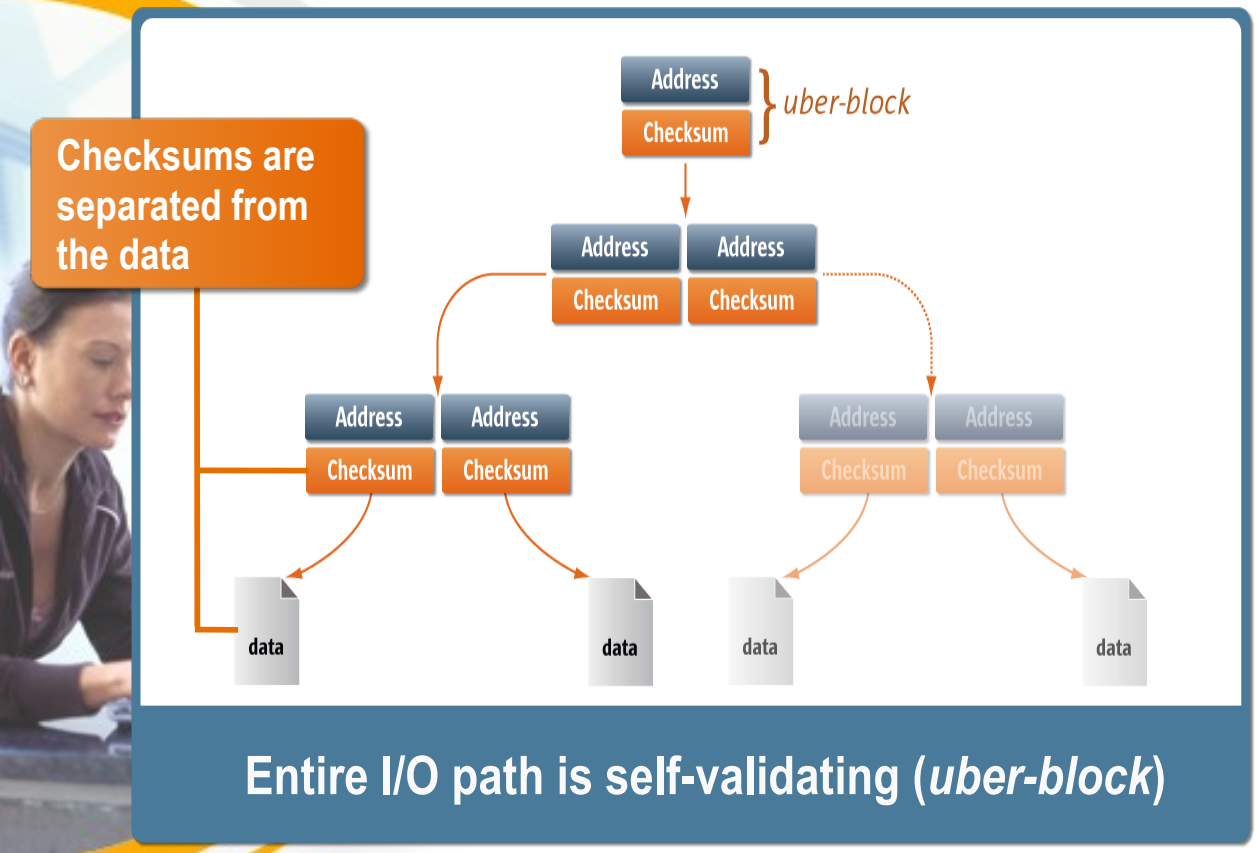
ZFS Data Integrity Model

- Copy-on-write, transactional design
- Everything is checksummed
- RAID-Z 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

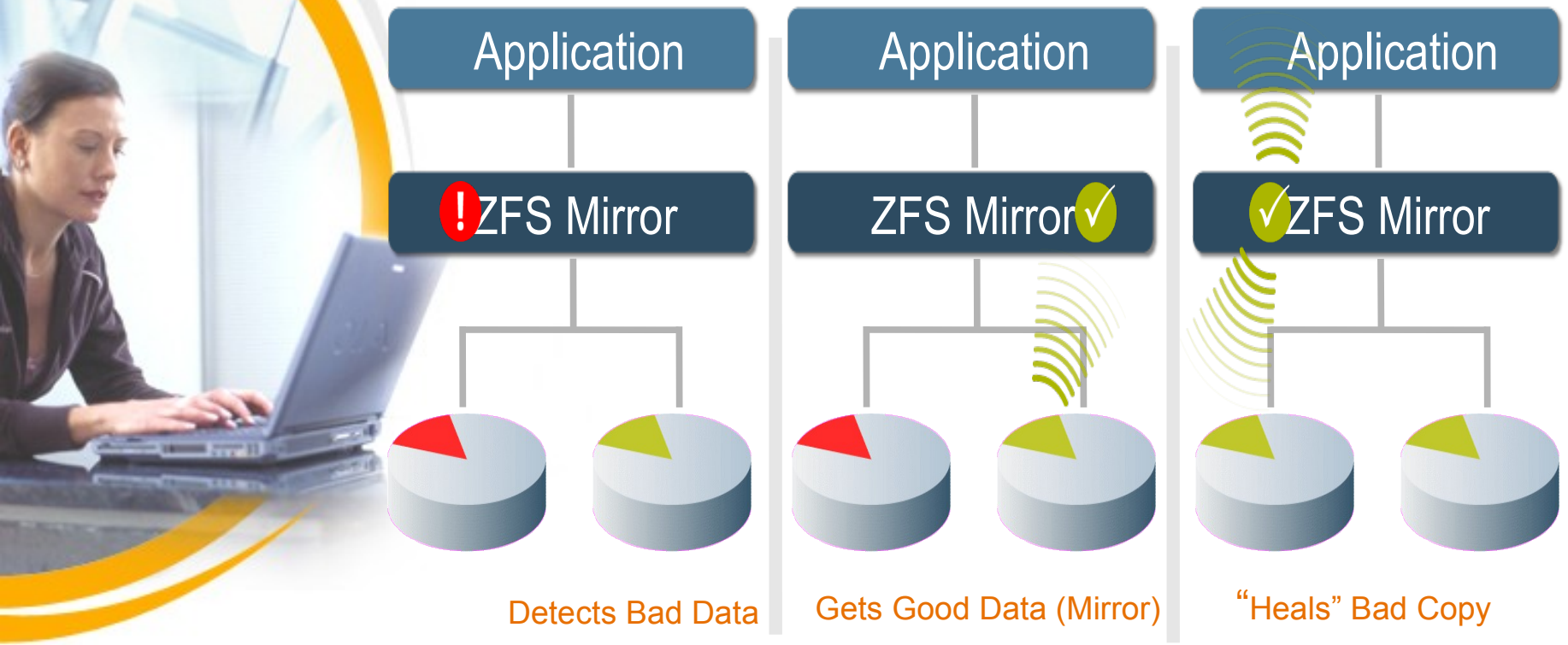
RAID-Z Protection

RAID-5 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

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

Immense Data Capacity

128-bit File System

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

Concurrent Everything

- EASIER

ADMINISTRATION

Easier Administration

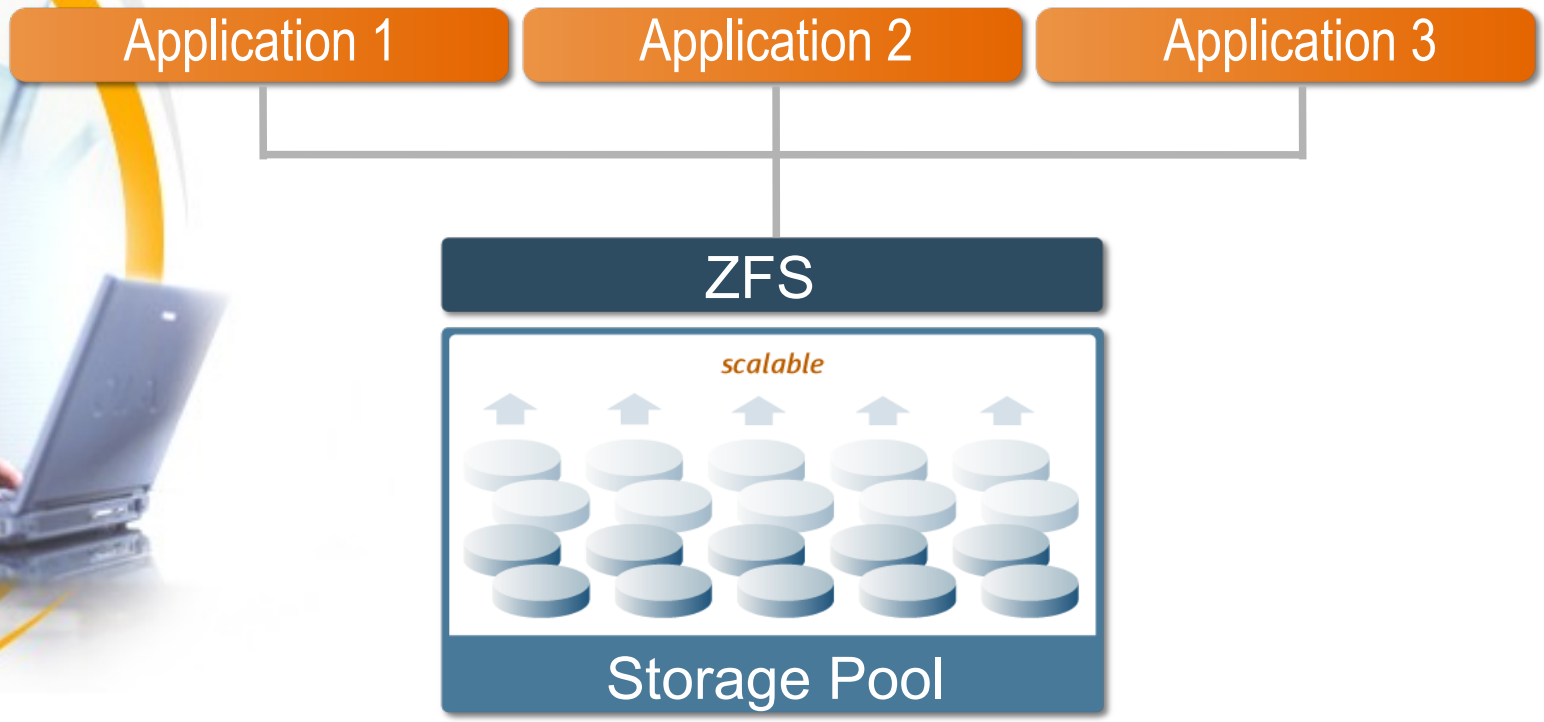
Pooled Storage Design makes for Easier Administration

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

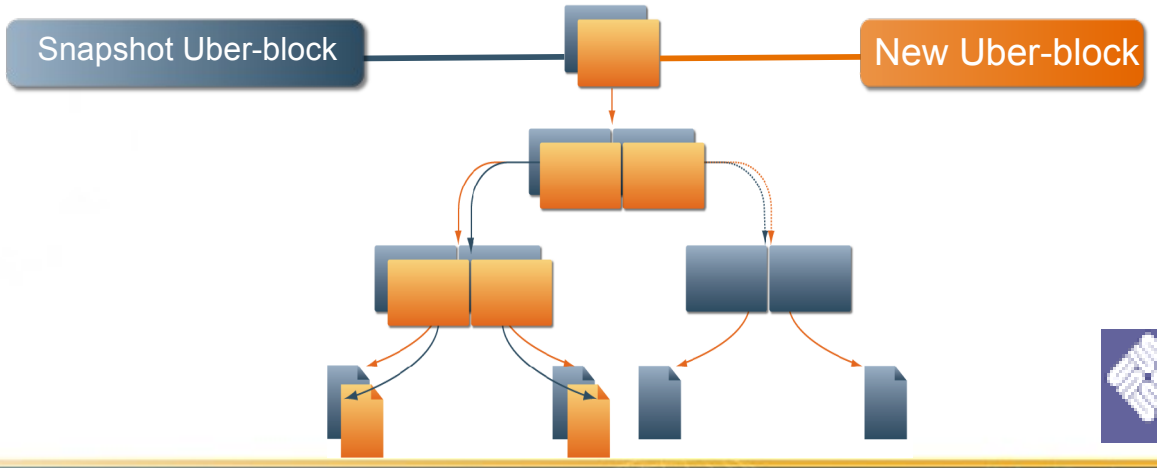


ZFS File systems are Hierarchical

- 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



ZFS Clones

Writable copy of a snapshot

Ideal for storing many private copies of shared data:

- Software installations
- Workspaces
- etc.

Example: Create a clone of Monday Status:

```
# zfs clone  ank/solaris@monday \
    tank/ws/yits/fix
```

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
```

Storage Pool Migration

“Adaptive Endian-ness”

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

Opposite “Endian” Systems

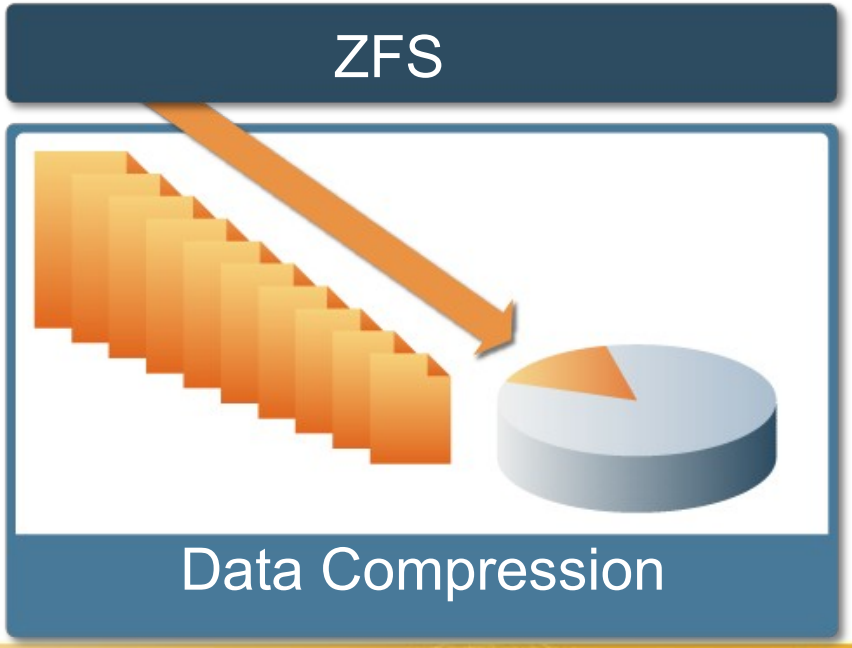
- Write and copy operations will eventually byte swap all data!

Config Data is Stored within the Data

- When the data moves, so does its config info

Data Compression

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



Data Security

- ACLs and Checksums

ACLs based on NFSv4 – NT style

- Full allow / deny semantics with inheritance
- Fine grained privilege control model (17 attributes)

The uber-block checksum can serve as a digital signature for the entire filesystem

- 256 bit, military grade checksum (SHA-256) available

- **BREATHTAKING**

PERFORMANCE

Architected for Speed

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

Cost and Source Code

ZFS is FREE

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

opensolaris™

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

otevřený 열린 مفتوح ανοικτό ಮುಕ್ತ libre
मुक्त öppen **open** פתוח 开放的
開 オープン ಮುಕ್ತ libero nyílt
放 வெளிப்படை açık livre offen
的 открытый

And for the Future

More Flexible

- Pool resize and device removal
- Booting / root file system

More Secure

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

More Reliable

- Fault Management Architecture Integration
- Hot spares
- DTrace providers

Questions?

e&m computing Ltd.

א.מ.ת

א.מ.ת מיחשוב בע"מ

Thank You