Distributed Storage: Parallel NFS and iSCSI with ZFS on OpenSolaris

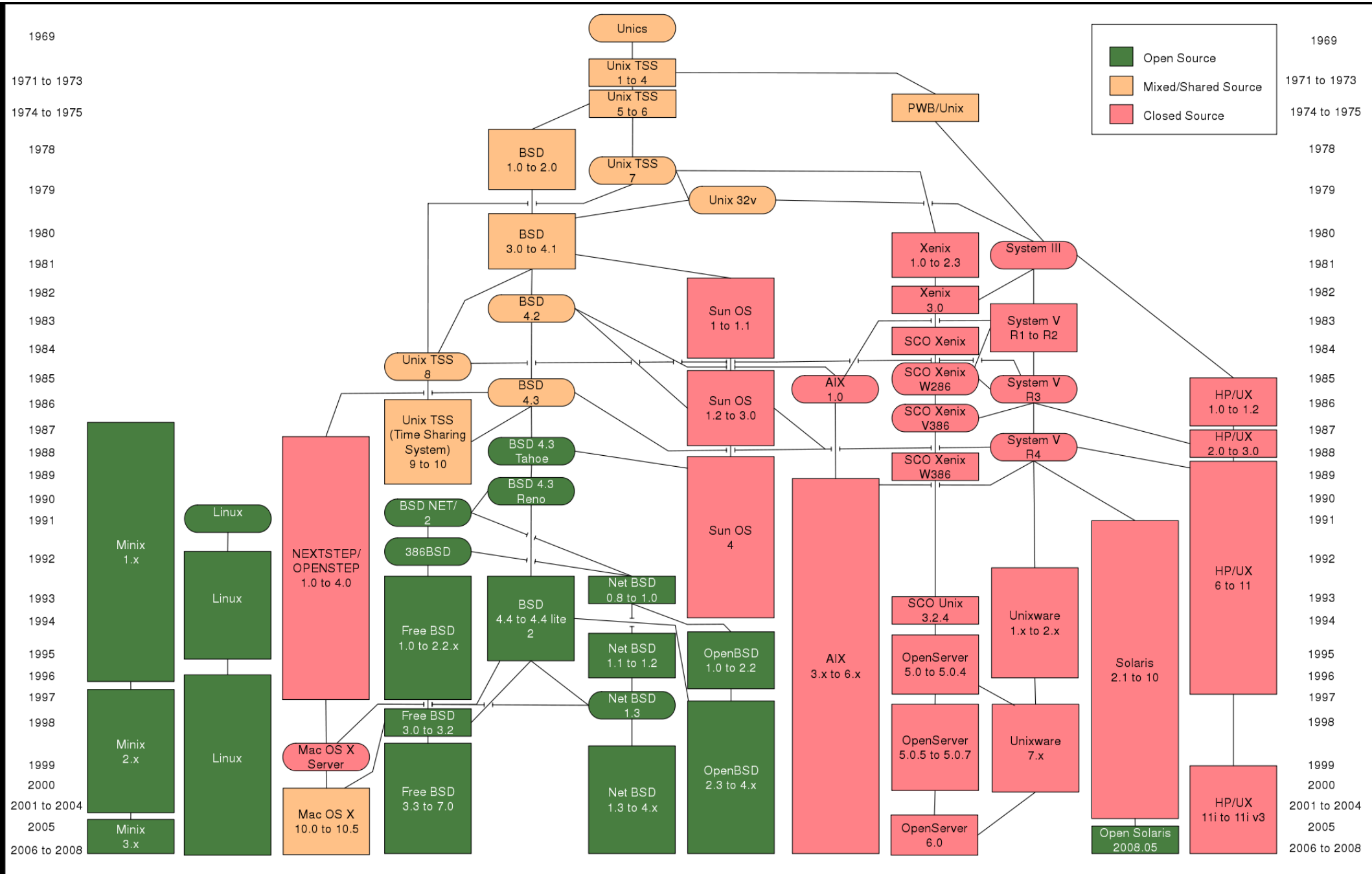
Kyriakos K. Skafas - kskafas@yahoo.gr

- Η διαρκώς αυξανόμενη παραγωγή και κατανάλωση πληροφοριών δημιουργεί την ανάγκη λόγω της μέχρι ενός ορίου κλιμάκωσης της χωρητικότητας και τις αξιοπιστίας σε μεμονωμένους εξυπηρετητές την κατανομή τους σε πολλαπλούς εξυπηρετητές.
- Το OpenSolaris με το σύστημα διαχείρισης αποθηκευτικών χώρων και αρχείων ZFS, με το σύστημα διαμοιρασμού αποθηκευτικών χώρων iSCSI και με το σύστημα διαμοιρασμού αρχείων σε δικτυακό περιβάλλον Parallel NFS προσφέρει μία αξιόπιστη λύση με τη χρήση αποκλειστικά Ελεύθερου και Ανοικτού Κώδικα Λογισμικού η οποία εξασφαλίζει αδιάληπτη διαθεσιμότητα, απόλυτη ακεραιότητα και ικανοποιητική ταχύτητα σε μεγάλα μεγέθη.

OpenSolaris

Ορισμοί

- UNIX
  - Genetic (both AT&T and BSD derivative)
  - Trademark or Branded
  - Functional (POSIX, SUS etc. compliant)
- Free and Open Source Software (Common Development and Distribution License)
- The only FOSS AT&T UNIX derivative


1969
1971 to 1973
1974 to 1975
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001 to 2004
2005
2006 to 2008
Open Source
Mixed/Shared Source
Closed Source
Unics
Unix TSS 1 to 4
Unix TSS 5 to 6
Unix TSS 7
PWB/Unix
Unix 32v
BSD 1.0 to 2.0
BSD 3.0 to 4.1
BSD 4.2
BSD 4.3
BSD 4.3 Tahoe
BSD 4.3 Reno
Unix TSS 8
Unix TSS (Time Sharing System) 9 to 10
BSD NET/ 2
386BSD
Free BSD 1.0 to 2.2.x
Free BSD 3.0 to 3.2
Free BSD 3.3 to 7.0
BSD 4.4 to 4.4 lite 2
Net BSD 0.8 to 1.0
Net BSD 1.1 to 1.2
Net BSD 1.3
Net BSD 1.3 to 4.x
OpenBSD 1.0 to 2.2
OpenBSD 2.3 to 4.x
Sun OS 1 to 1.1
Sun OS 1.2 to 3.0
Sun OS 4
AIX 1.0
AIX 3.x to 6.x
Xenix 1.0 to 2.3
Xenix 3.0
SCO Xenix
SCO Xenix W286
SCO Xenix V386
SCO Xenix W386
SCO Unix 3.2.4
OpenServer 5.0 to 5.0.4
OpenServer 5.0.5 to 5.0.7
OpenServer 6.0
System III
System V R1 to R2
System V R3
System V R4
Unixware 1.x to 2.x
Unixware 7.x
Solaris 2.1 to 10
Open Solaris 2008.05
HP/UX 1.0 to 1.2
HP/UX 2.0 to 3.0
HP/UX 6 to 11
HP/UX 11i to 11i v3
Minix 1.x
Minix 2.x
Minix 3.x
Linux
Linux
Linux
NEXTSTEP/ OPENSTEP 1.0 to 4.0
Mac OS X Server
Mac OS X 10.0 to 10.5

OpenSolaris

Πηγές

- http://en.wikipedia.org/wiki/Solaris_%28operating_system%29
- http://en.wikipedia.org/wiki/Opensolaris
- http://hub.opensolaris.org/bin/view/Main/
- http://www.opensolaris.com/
- http://www.opensolaris.com/learn/
- http://www.sun.com/cddl/cddl.html

ZFS

Δυνατότητες

- Storage pools (disks, partitions, slices, files...)
- Capacity (128bit and 64bit limits)
- Copy-on-write
- Snapshots (non writeable) and clones (writeable)
- Dynamic striping
- Variable block sizes
- Lightweight filesystem creation
- Cache management
- Adaptive endianness
- De-duplication

ZFS

Δυνατότητες (συνέχεια)

- I/O priority with deadline scheduling
- I/O sorting and aggregation
- Multiple independent prefetch streams
- Parallel, constant-time directory operations
- End-to-end checksumming
- Transparent filesystem compression
- Intelligent scrubbing and resilvering
- Load and space usage sharing between disks in the pool
- Ditto blocks
- Per-user and per-group quotas support

ZFS

Μελλοντικές Δυνατότητες

- Transparent filesystem encryption
- Resizing pools
- Defragmentation
- Re-compression

ZFS

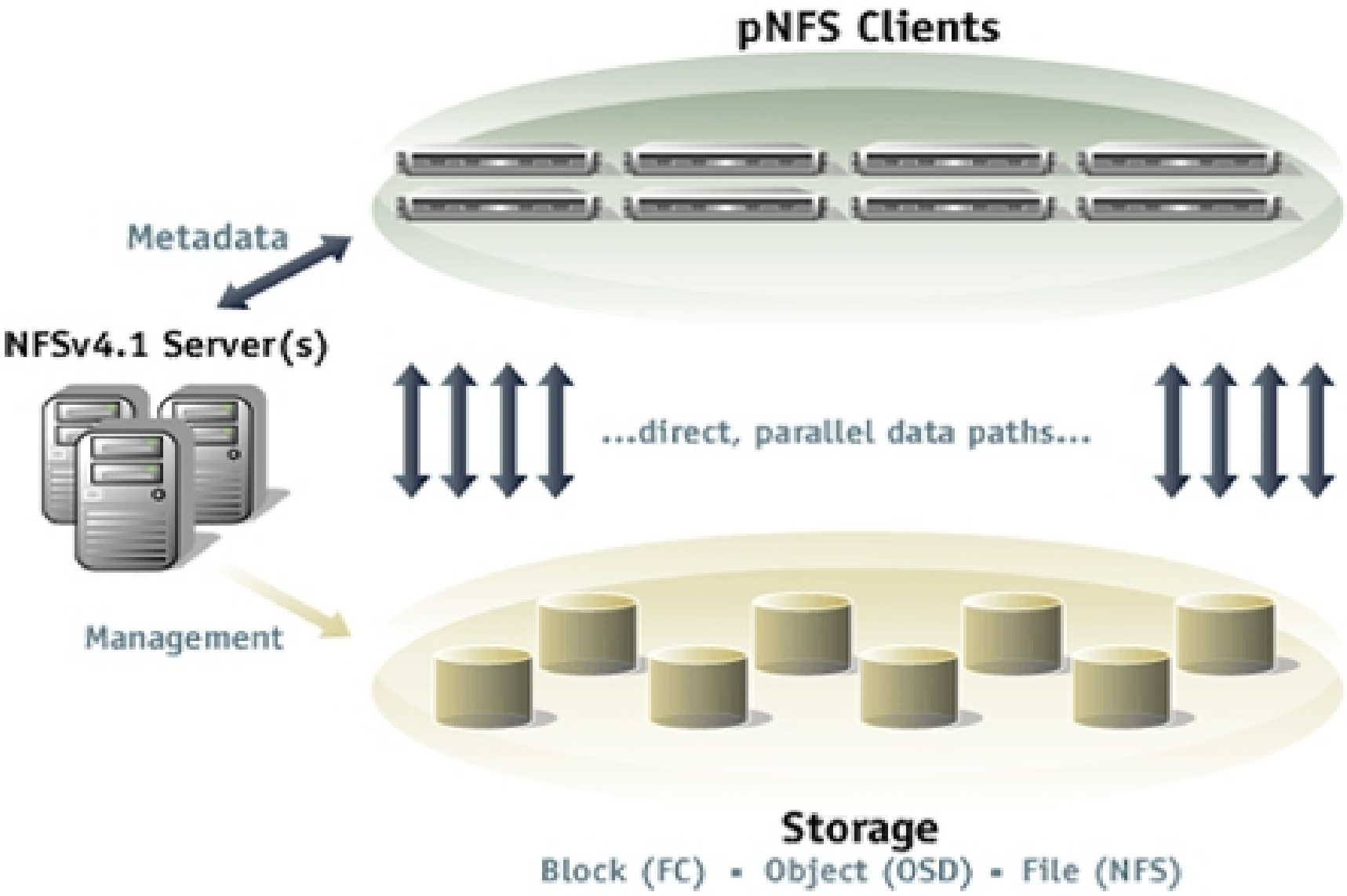
Πηγές

- http://en.wikipedia.org/wiki/ZFS
- http://www.opensolaris.com/learn/features/storage/
- http://www.opensolaris.com/learn/features/solariszfs.pdf

NFS/pNFS

Ορισμοί

- Network File System (clients, servers)
  - NFS v2
  - NFS v3
  - NFS v4
- Parallel Network File System (clients, data servers, meta-data servers)
  - NFS 4.1 (backwards compatible)


pNFS Clients
Metadata
NFSv4.1 Server(s)
...direct, parallel data paths...
Management
Storage
Block (FC) • Object (OSD) • File (NFS)

NFS/pNFS

Πηγές

- http://en.wikipedia.org/wiki/Network_File_System_%28protocol%29
- http://www.pnfs.com/
- http://wikis.sun.com/display/NFS/Latest+-+pNFS+Admin+Documentation
- http://wikis.sun.com/display/NFS/pNFS+Release+Notes

iSCSI

Ορισμοί

- SCSI-over-IP
- iSCSI clients are called Initiators
- iSCSI servers are called Targets

iSCSI

Common Multiprotocol SCSI TARget (COMSTAR)

- In-kernel
- Target and Initiator
- LUN masking and mapping
- Multipathing
- Parallel transfers

iSCSI

Πηγές

- http://en.wikipedia.org/wiki/Iscsi
- http://wikis.sun.com/display/OpenSolarisInfo/comstar+Administration
- http://hub.opensolaris.org/bin/view/Project+comstar/

Installation and Service Management

- Install

```
# pkg install storage-server SUNWiscsit SUNWiscsi
```

- Disable legacy Services

```
# svcadm disable svc:/system/iscsitgt:default
```

- Start Services

```
# svcadm enable svc:/system/stmf:default
# svcadm enable svc:/network/iscsi/target:default
```

- Check Services

```
# svcs svc:/system/stmf:default
# svcs svc:/network/iscsi/target:default
```

- Check State

```
# stmfadm list-state
```

Type of Logical Unit

- Decide on type of LU
  - File-based
  - Thin-Provisioned
  - Disk/Disk Partition
  - ZFS Volume

Target: Create a Logical Unit

- Create ZFS Volume

```
# zfs create -b <blocksize> -V 16G dpool/dvol
```

- Create LU

```
# sbdadm create-lu /dev/zvol/rdsk/dpool/dvol
```

- Check LU

```
# sbdadm list-lu
```

Target: Serve Logical Unit

- Get Global Unique Identification (GUID)

```
# sbdadm list-lu
```

- Add view

```
# stmfadm add-view <GUID>
```

- Check view

```
# stmfadm list-view -l <GUID>
```

- Create Target Portal Group and then Target

```
# itadm create-tpg <interface> <address>
```

```
# itadm create-target -t <interface>
```

- Check Targets

```
# itadm list-target -v
```

Initiator: Target Discovery

- Target discovery

```
# iscsiadm add discovery-address <target address>
# iscsiadm modify discovery --sendtargets enable
```

- Check discovery method

```
# iscsiadm list discovery
```

- Check discovered targets

```
# iscsiadm list target
```

- Create device files

```
# devfsadm -i iscsi
```

- List available LU's

```
# format
```

ZFS

- Create pool

```
# zpool create idvol mirror <disk0> <disk1>
```

- Create file system

```
# zfs create idvol/dfs
```

Monitoring

- Process

  `prstat, top...`

- File systems

  `fsstat, zfs stat...`

Performance

- Enable or disable write cache on the target:

```
# stmfadm modify-lu -p wcd=false/true <LU>
```

- Disable Nagle's Algorithm on the initiator:

```
/kernel/drv/iscsi.conf:

...

tcp-nodelay=1;

...
```

- Generic Interface, IP, TCP tunables...
  - Jumbo Frames
  - Large send and receive window...

Ευχαριστίες

- Διοργανωτές FOSSCOM 2010 @ Thessaloniki
  - Ρούζη Στέλλα
  - Μπαχαράκη Χρήστο
  - Χουτουρίδη Χρήστο
- Κοινότητα ΕΛ/ΛΑΚ
- Κοινότητα HELLUG
- Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
- ΕΣΑΣ! ;-)

Ερωτήσεις; Απορίες;