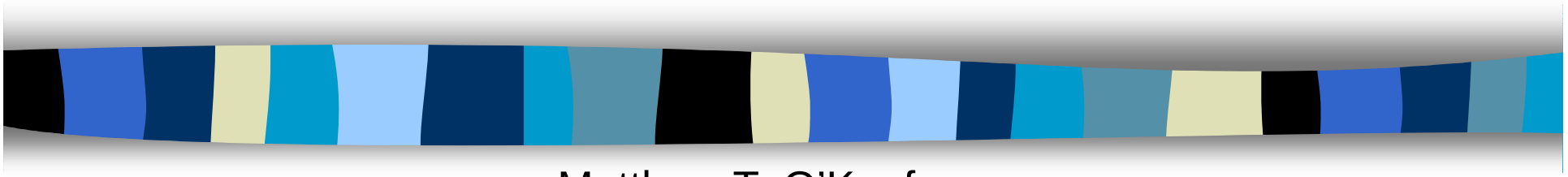
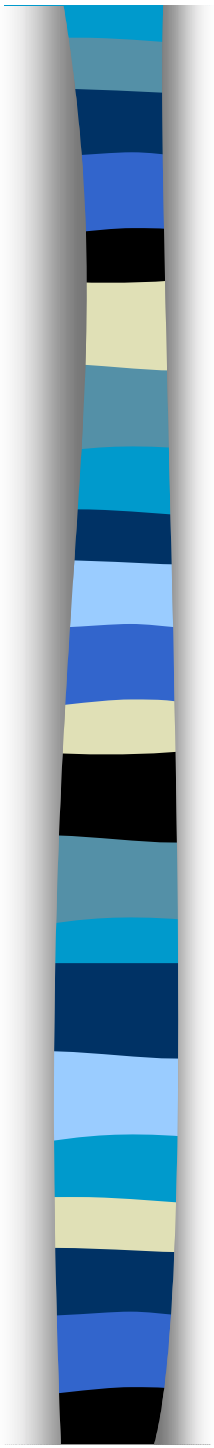
Open Source Storage Management in Linux



Matthew T. O'Keefe
University of Minnesota
and
Sistina Software, Inc.

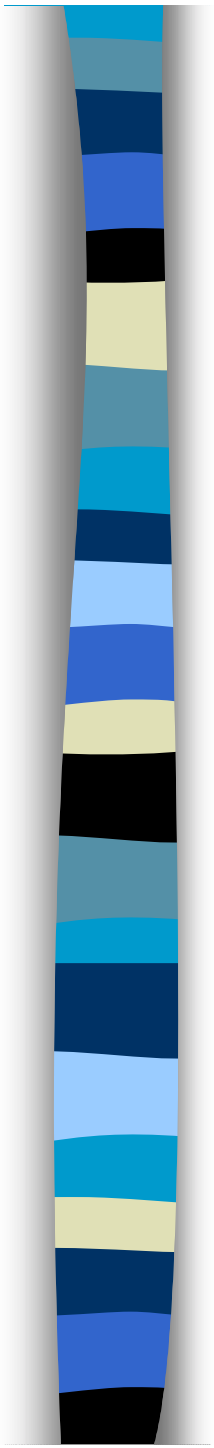
University of Minnesota

- Parallel Computer Systems Laboratory (PCSL)
- http://www.borg.umn.edu
- Two areas of expertise:
  - parallel simulation software and development environments
    - time-domain electromagnetics
    - fluid dynamics (in particular, ocean modeling)
  - storage management software
    - global file systems and volume management
    - secure file system
    - fibre channel and storage area networking



Parallel Computer Systems Laboratory

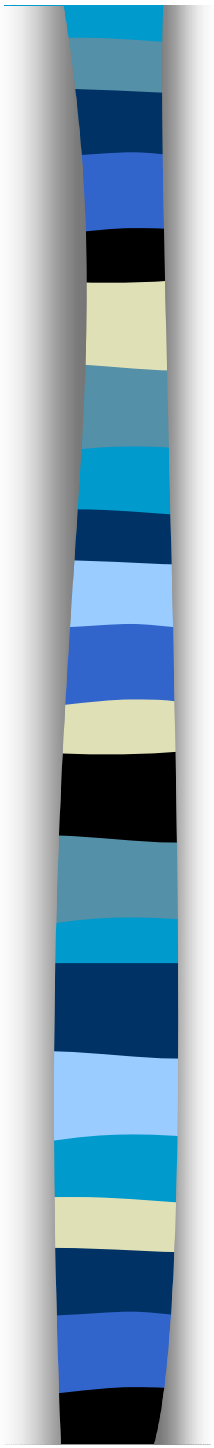
- 15 undergraduates, 10 graduate students, 1 part-time staffer
- Funding sources include
  - federal agencies (ONR)
  - contracts with Companies (STK, Brocade) + equipment
  - annual sponsorship fees: sponsors include
    - Brocade
    - StorageTek
    - Seagate
    - EMC
    - SGI
    - Veritas Software



Parallel Computer Systems Laboratory

- Laboratory goal and vision is very simple:

Write great software and give it away.

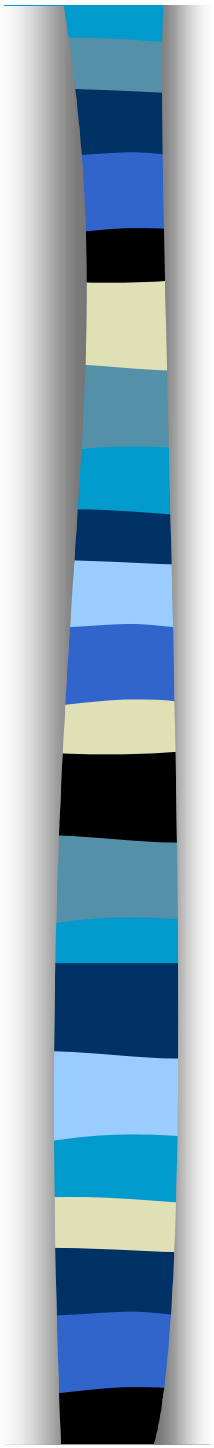


Parallel Computer Systems Laboratory

- Good things that fall out of this simple goal:
  - first and foremost, *educating the next generation* of computer system designers and architects
    - this requires years of implementation experience
    - students do real implementation work
  - *setting open industry standards* through high-quality open source software
    - global file systems
    - secure file systems
    - time-domain electromagnetics and fluid dynamics
    - storage area networking software

Simulation and Storage

- Our parallel simulations generate huge amounts of data, which we need to post-process and archive
- We saw Linux and SANs as a solution to this problem, but we needed a cluster file system to do it right



Linux Open Source Storage Technology

- Device drivers and interfaces
- Logical volume managers
- Software RAID
- Local File Systems
- Cluster File Systems
- Distributed File Systems
- Backup and Restore
- High Availability
- Note: I will only talk about open source code

Device Drivers and Interfaces

- Parallel SCSI (Adaptec, etc.)
  - Adaptec quote: “the drivers written by the Linux community for our SCSI cards are better than the drivers written by our own programmers”
  - mid-layer needs work (error reporting and recovery)
- Fibre Channel
  - Qlogic FC adapters, Fabric and Loop support
- USB, Firewire, etc

Logical Volume Managers

- Logical Volume Manager (H. Mauelshagen)
- MD (Multiple Devices) Software RAID driver
- Pool Volume Manager — for clusters of machines sharing disks

Software RAID

- MD Driver supports RAID O, 1, 3, 4, 5
- Recovers on disk failure
- Used by Linux community as a cheap form of RAID
- Inexpensive is VERY important to Linux developers
  - a big part of Linux's strength is that it is inexpensive
  - inexpensive and good enough usually wins out over expensive and full-featured in technology competition (unless there is no competition)

Local File Systems

- ext2fs: traditional UNIX file system, very fast, unsafe
- ext3fs: journaled version of ext2fs, very fast, safe
- reiserfs: journaled, uses B-trees
- SGI's XFS: journaled, B-trees, delayed allocation — being ported from IRIX

Cluster File Systems

- Global File System (GFS): allows machines on a storage network to share disks
- A SAN File System
- Symmetric, simple cluster membership services
- Beta release with journaling and recovery
- See full talk on Wednesday by K. Preslan

Distributed File Systems

- Both CIFS and NFS supported: Linux becoming popular with storage appliance vendors
- Samba (supports SMB clients, makes Linux look like an NT server)
  - better support for SMB than Microsoft's own server platforms
- NFS client and server support
- Coda from CMU
- Intermezzo: exciting new development from Peter Braam

Backup and Restore

- Amanda (Advanced Maryland Automated Network Disk Archiver) program developed at U. Maryland
- Multiple UNIX clients can back up to single server which has the tape device
  - multiple clients write backup data in parallel (using dump or GNU tar) to a “holding disk”
  - backup images in the holding disk are aggregated to create a single large write stream to the Amanda tape device
  - use SAMBA to support Windows clients
- Limitations:
  - can’t write multiple data streams to tape
  - can’t write single dump image across multiple tapes

High Availability

- www.linux-ha.org
- Several on-going efforts
  - Tweediecluster — like Vaxcluster only better
  - SGI's FailSafe being ported
  - Lots of application specific HA work (web serving)

Linux and Storage Management

- Users who make the switch to Linux will benefit from the commoditization of storage management software
  - volume management, cluster file systems, HA
- These users will have access to source code, can maintain code themselves, modify it as needed for their own purposes

Open Source Licenses

- Open Source definition
  - no royalties, no warranties
  - access to source code
  - right to modify and re-distribute without restriction
  - rights given to all, must not discriminate
- BSD License — least restrictive
- GNU Public License — access to source guaranteed
- Mozilla Public License — special rights to originator

Open Source Storage Management Software

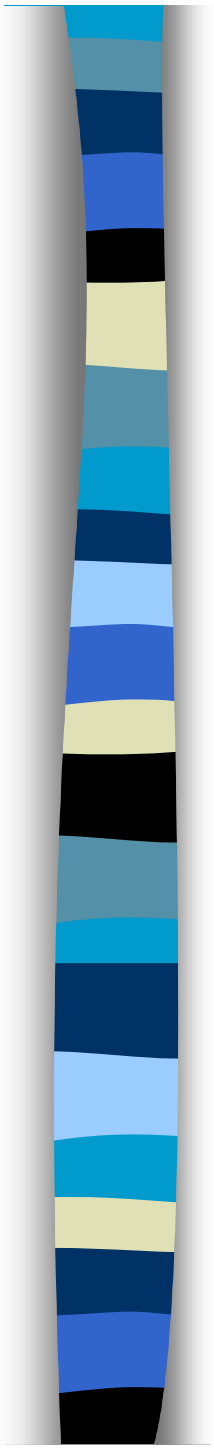
- Use the Internet model for open source storage management software development
  - “rough consensus and running code”
  - open source ALL the way
    - Bind, Perl, *BSD + Linux, Apache, TCP/IP stack
- Storage area networking software must be done the same way
  - simple protocols and shared source code (GFS, Pool VM)
  - global file systems and SAN device drivers must be developed in open source
    - to build a SAN infrastructure requires shared source code, just like the Internet

Linux Origins

- Basically a hobby that got WAY out of control…
- Developed by Linus Torvalds and a loosely-organized band of hackers on the Internet
- Fastest growing enterprise OS by a wide margin
- Uses GPL (no royalties, access to source code rights is perpetuated)
  - no royalties is important to peace among developers
- Attacking desktop, servers, and embedded markets

Linux Development Model

- Best explained by Eric Raymond in his paper “The Cathedral and the Bazaar”: www.opensource.org
- Refutes Brooks’ mythical man-month principle that adding more programmers makes the schedule later
  - debugging code in parallel is possible
  - writing loadable kernel modules in parallel is possible
- Raymond paper “Homesteading the Noosphere” describes hacker open source culture as similar to gift culture



Why Open Source Yields Better Systems

- Better architectures
  - very intense peer review
    - read the Linux kernel mailing lists: it is one long-running FLAME-FEST! (nothing personal, of course :)
  - open source licenses encourage others to contribute
    - GPL versus BSD
- Better code
  - very intense peer review
  - many alternative implementations
    - there are 4 journaled file system in Linux right now!
    - 3 network block drivers
    - multiple efforts to improve SCSI mid-layer
    - etc.

Why Open Source Yields Better Systems

- Better Debugging
  - very intense, parallel debugging effort (contrast closed- versus open-source)
    - if your beta users have the source they can find the bugs for you
    - with Linux its not just the source code, but a very detailed archival record of how and why a particular piece of code was written the way it was, plus all those O'Reilly books :)
  - "with enough eyeballs, all bugs are shallow"
  - people are willing to use raw, pre-beta open source code in because
    - they aren't paying for it directly (paying with their time)
    - they become part of the development process, which makes for a very strong interaction with key early adopters

Why Open Source Yields Better Systems

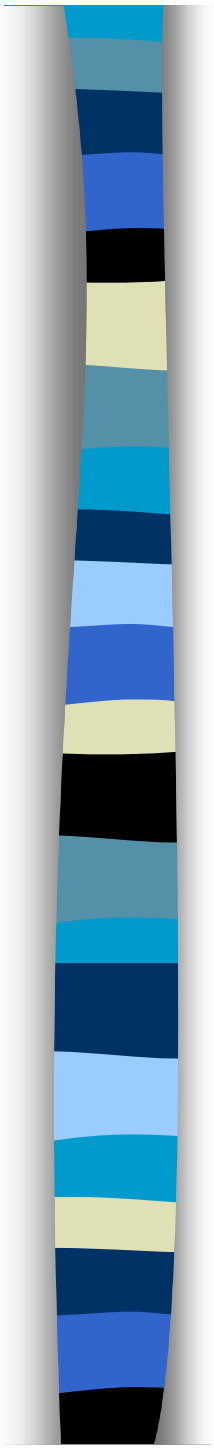
- Better software maintenance
  - open source code does not go away when “the company” goes out of business
  - usually, if it is an important and widely-used piece of code is no longer maintained by the original developer someone else will pick up the task of maintaining the code
  - the software can be customized to fit the needs of the user
    - very important to demanding users
    - early adopters can help each other and share customization costs
  - maintenance costs can be shared among many users

Open Source Economics

- What happens to the value of a particular piece of software when the vendor developing it
  - goes out of business or
  - stops supporting the product?
- When maintenance and future support for software is dropped by vendor, the sale value of the software drops to zero
  - manufacturing versus service view of software

Open Source Economics

- Linux business models
  - Linux software distribution (Red Hat, Caldera, Suse)
    - sell media, manual, brand recognition
  - Linux hardware OEMs (VA Linux, Alta Technology)
  - Linux support — Linuxcare (KP-funded) just does generic Linux support
  - Server appliances (Cobalt Networks, many others)
  - Sistina model: we are an arms merchant — we sell technology and services to users and suppliers
- Linux and the stock market
    - VA Linux, Red Hat, Cobalt have done well (maybe too well)

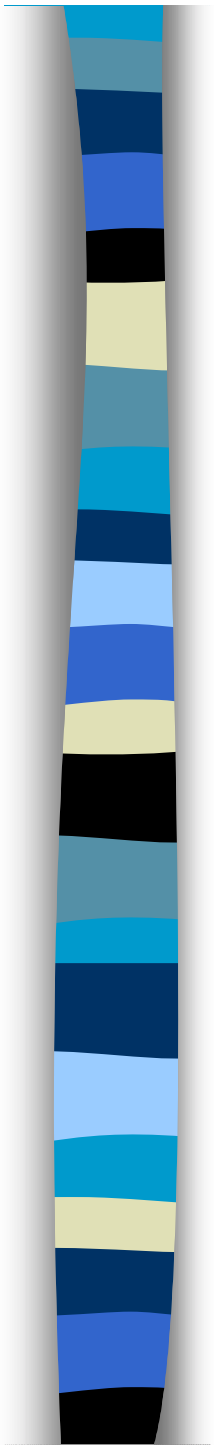


Open Source Economics

- Excitement around Linux due to the chance it will become the standard software platform for servers and server appliances
  - open-source is the next wave in computing
  - the “next big thing” is usuallly a lot less expensive but less capable initially than the current “big thing”
  - but soon replaces the latter due to volume and huge infrastructure investments: feature equivalence quickly achieved
  - Innovator’s Dilemma

Predictions

- By 2002 Linux will have completely eclipsed every other enterprise OS
  - all other closed-source UNIXen are already legacy systems (including Solaris, IRIX, AIX, Digital UNIX, HP-UX, and Windows NT)
- UNIX OEMs like HP, Sun, SGI will be overtaken by fast, nimble collection of Linux hardware and software vendors
  - their cost structures are fundamentally wrong: Linux model of shared development means the Linux cost structure is lower
  - Linux runs on all hardware platforms, which will result in fierce competition, lower costs, and huge price/performance advantages



Resources

- Linux Storage Management
  - http://www.globalfilesystem.org
- Press releases and general news on Linux
  - http://www.linuxtoday.com
- Real code and kernel mailing list
  - http://www.kernelnotes.org
- Gossip
  - http://www.slashdot.org