Construction and Operation of a MOSIX analysis cluster

A.Korn,C.Paus
Massachusetts Institute of Technology
akorn@mit.edu

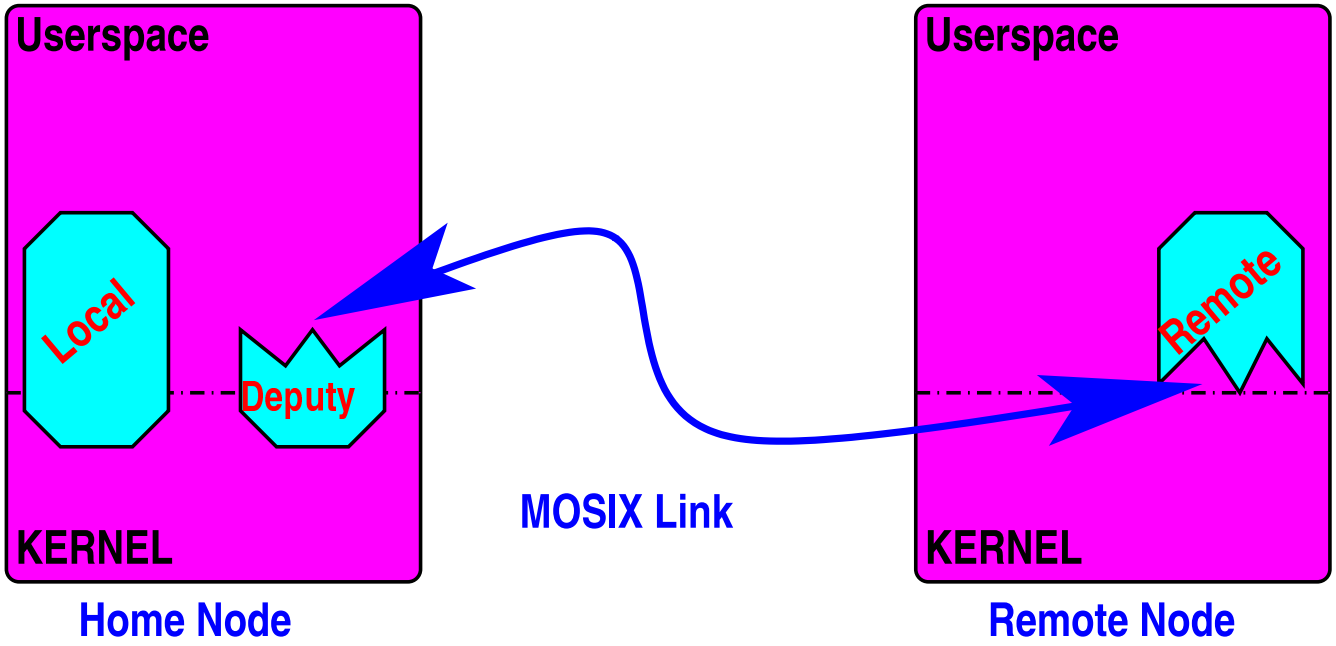
October 21, 2002

Motivation

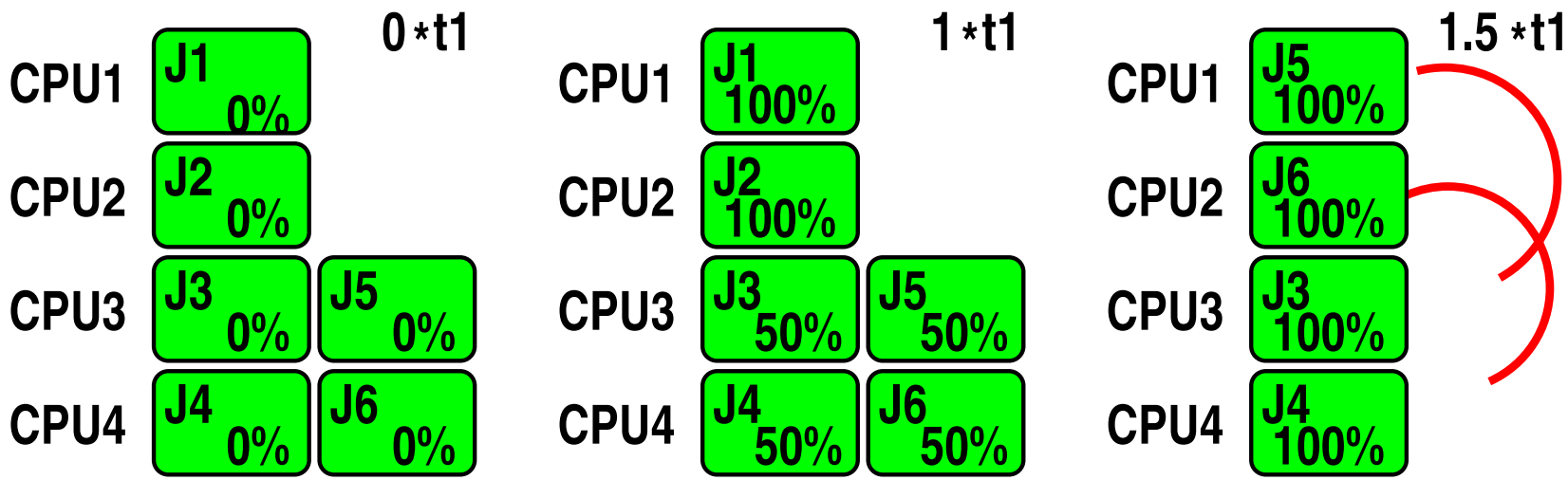
- Needed a group computing facility for CDF analysis
- Maximize compute power and user convinience
- Minimize cost
- Use commodity components (standard PC's)
- Allow for interactive users
- Single System Image appearence (like a large standard SMP machine)

MOSIX Kernel

- **M**ulti-PC **O**perating **S**ystem for Un**IX** (www.mosix.org)
- Linux kernel patch developed by Hebrew University of Jerusalem
- Transparent process migration
- Dynamic load balancing
- Home node concept



Dynamic load balancing

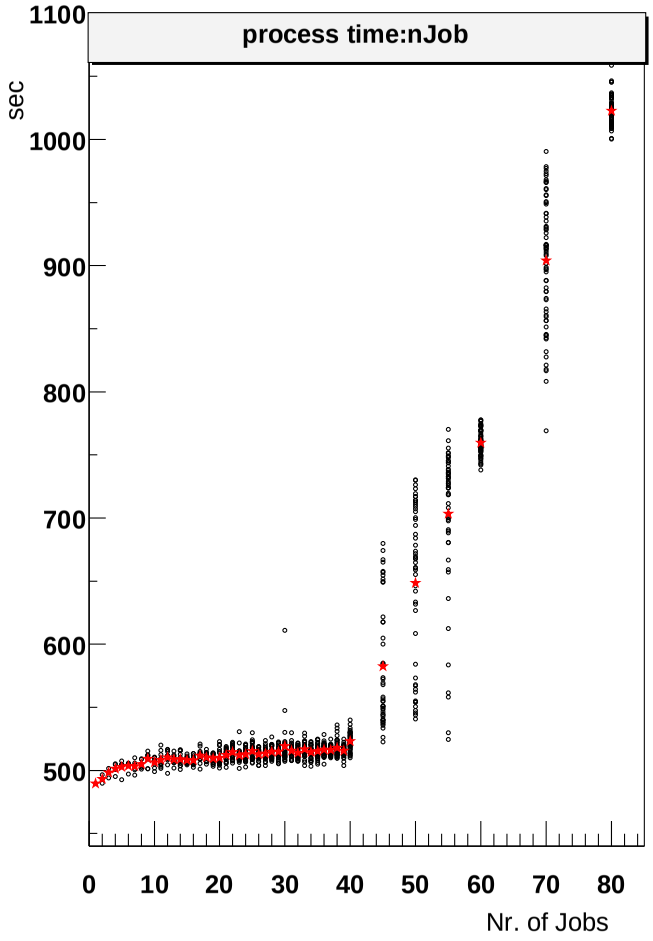


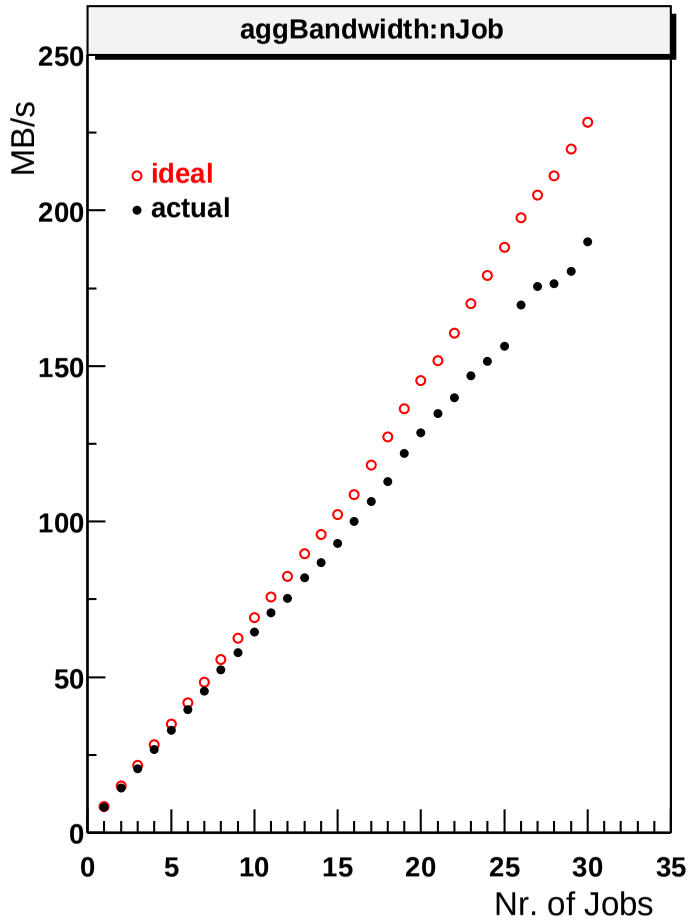
$\Longrightarrow$ jobs migrate to free nodes

DFSA/MFS

- **D**irect **F**ile**S**ystem **A**ccess (DFSA) process moves to data
- Needs special fs features
  $\Longrightarrow$ use MFS (Mosix File System)
- Disks distributed over all nodes
- Each disk visible on all nodes
- e.g. access /mfs/6/disk3 etc.
- MFS speed close to NFS for nonlocal data access

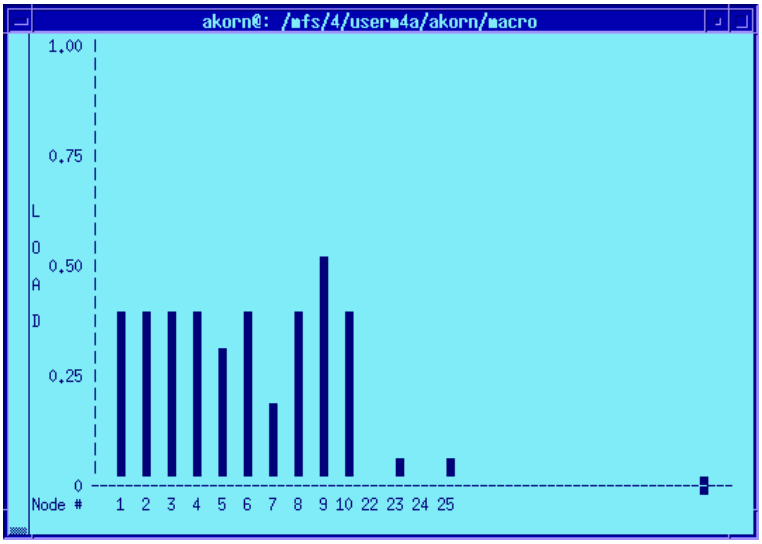
Test system Benchmarks





Tools

- Monitoring
  - mon



  - qps

- Migration characteristics
  - mosrun (change job behaviour)
  - mosctl (change node behaviour)

- User convenience
  - diskspace

qps@b0pcmit13

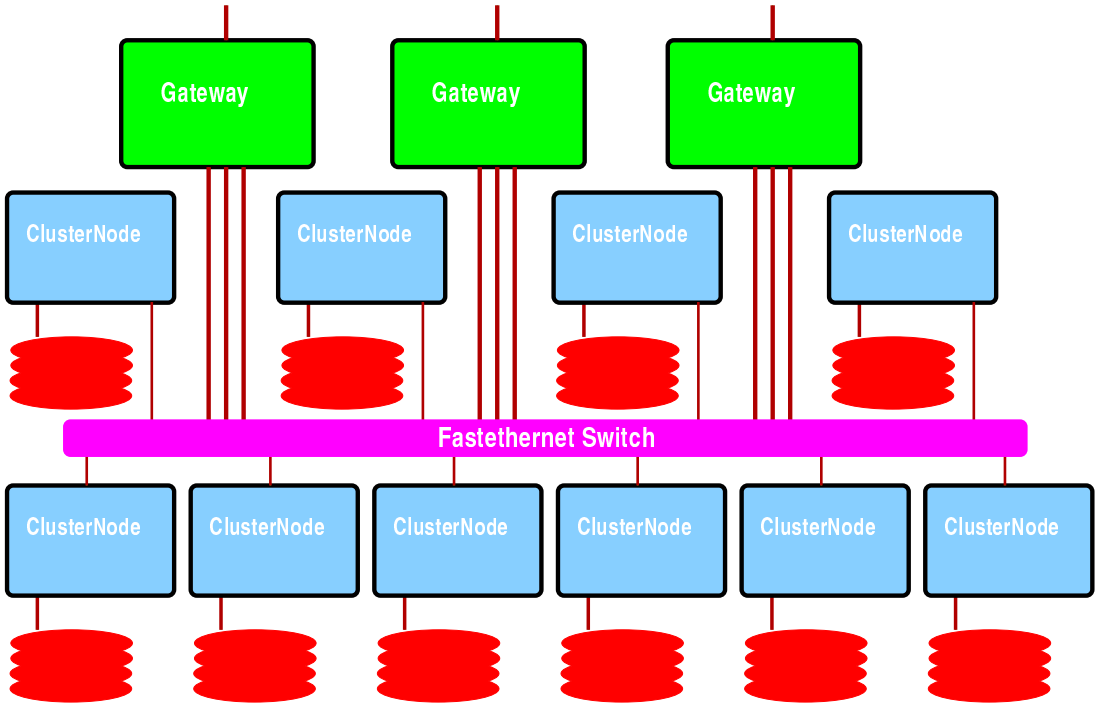
Process Signal View Command Options Help

cpu: user nice sys idle | mem: used buff cach free | swap | up 1 day, 19:48

Linear Tree Update

PID	TTY	USER	NICE	MIGR	NMIGS	NOMOVE	SIZE	RSS	STAT	%CPU	START	TIME	CPU	CMDLINE
15818	pts/3	anikeev	0	>4	2		73632	73632	S	99.04	15:11	7:40	0	./bin/Linux2-KCC_4_0/mu.exe simTCMD.tcl
15846	pts/3	anikeev	0	>8	3		72352	72352	S	98.08	15:12	6:49	0	./bin/Linux2-KCC_4_0/mu.exe simTCMD.tcl
15835	pts/3	anikeev	0	>7	3		72032	72032	S	97.89	15:12	6:47	0	./bin/Linux2-KCC_4_0/mu.exe simTCMD.tcl
15833	pts/3	anikeev	0	>6	2		72992	72992	S	97.12	15:11	7:07	0	./bin/Linux2-KCC_4_0/mu.exe simTCMD.tcl
15820	pts/3	anikeev	0	>5	4		71628	71628	S	96.15	15:11	7:03	0	./bin/Linux2-KCC_4_0/mu.exe simTCMD.tcl
15816	pts/3	anikeev	0	>2	4		71548	71548	S	87.26	15:11	6:46	0	./bin/Linux2-KCC_4_0/mu.exe simTCMD.tcl
15831	pts/3	anikeev	0	>1	4		71500	71500	S	82.53	15:11	6:26	0	./bin/Linux2-KCC_4_0/mu.exe simTCMD.tcl
807	-	root	0	-	0	monkey, mmap_dev, direct_io	21632	21632	S	2.11	Oct 14	7:59	0	/etc/X11/X -auth /var/gdm/:0.Xauth :0
15897	pts/5	akorn	0	-	0		4356	4356	R	2.11	15:13	1.36s	1	/mfs/22/bin/qps
1578	pts/3	anikeev	0	-	0		4512	4512	S	1.93	Oct 14	35:52	1	/mfs/22/bin/qps
772	-	root	20	-	0	daemon, clone_vm, rt_sched	0	0	SWN	1.15	Oct 14	12:40	1	(infodaemon)
821	-	bauerg	0	-	0		1280	1280	S	0.19	Oct 14	7.90s	0	/usr/bin/X11/fvwm2
1	-	root	0	-	0	init_proc	476	476	S	0.00	Oct 14	0:11	0	init [5]
2	-	root	0	-	0	clone_vm	0	0	SW	0.00	Oct 14	5.51s	1	(kflushd)
3	-	root	0	-	0	clone_vm	0	0	SW	0.00	Oct 14	2.11s	1	(kupdate)
4	-	root	0	-	0	clone_vm	0	0	SW	0.00	Oct 14	0:12	1	(kswapd)
5	-	root	0	-	0	clone_vm	0	0	SW	0.00	Oct 14	0.00s	0	(keventd)
7	-	root	20	-	0	daemon, rt_sched	0	0	SWN	0.00	Oct 14	0.50s	1	(mfs_main_server)
8	-	root	0	-	0	daemon, clone_vm	0	0	SW	0.00	Oct 14	0.01s	1	(migdaemon)
9	-	root	0	-	0	daemon, clone_vm	0	0	SW	0.00	Oct 14	0.16s	0	(mfs_gc)
462	-	bin	0	-	0		536	536	S	0.00	Oct 14	2.28s	1	portmap
488	-	root	0	-	0		732	732	S	0.00	Oct 14	0.01s	0	rpc.statd
528	-	root	0	-	0	clone_vm	0	0	SW	0.00	Oct 14	8.08s	0	(rpciod)
529	-	root	0	-	0	clone_vm	0	0	SW	0.00	Oct 14	0.00s	0	(lockd)
548	-	root	0	-	0		524	524	S	0.00	Oct 14	1.69s	1	syslogd -m 0
559	-	root	0	-	0		804	804	S	0.00	Oct 14	0.11s	0	klogd
575	-	root+	0	-	0		496	496	S	0.00	Oct 14	0.01s	1	/usr/sbin/atd
591	-	root	0	-	0		620	620	S	0.00	Oct 14	0.31s	0	crond
611	-	root	0	-	0		484	484	S	0.00	Oct 14	0.02s	0	inetd
620	-	root	0	-	0		700	700	S	0.00	Oct 14	3.60s	1	/usr/sbin/sshd
638	-	root	0	-	0	rt_sched	1364	1364	S	0.00	Oct 14	0:12	0	xntpd -A
659	-	root	0	-	0	clone_vm	876	876	S	0.00	Oct 14	0.00s	0	ypbind
661	-	root	0	-	0	clone_vm	876	876	S	0.00	Oct 14	9.95s	1	ypbind
662	-	root	0	-	0	clone_vm	876	876	S	0.00	Oct 14	0.00s	0	ypbind
663	-	root	0	-	0	clone_vm	876	876	S	0.00	Oct 14	1.62s	1	ypbind
705	-	root	0	-	0		1204	1204	S	0.00	Oct 14	1.37s	1	sendmail: accepting connections on port 25
722	-	root	0	-	0		460	460	S	0.00	Oct 14	0:19	0	gpm -t ps/2
763	-	xfs	0	-	0		2288	2288	S	0.00	Oct 14	2.02s	0	xfs -droppriv -daemon -port -1
791	1	root	0	-	0		412	412	S	0.00	Oct 14	0.01s	1	/sbin/mingetty tty1
792	2	root	0	-	0		412	412	S	0.00	Oct 14	0.00s	0	/sbin/mingetty tty2
793	3	root	0	-	0		412	412	S	0.00	Oct 14	0.00s	0	/sbin/mingetty tty3
794	4	root	0	-	0		412	412	S	0.00	Oct 14	0.00s	0	/sbin/mingetty tty4

Current Configuration



- 10 Dual Pentium III 1GHz
- 6 Gateways (Dual Pentium III)
- ca. 10000 MIPS
- 2.5 Tb Harddisk space
- Total cost: $ 30k (Spring 2001)

Hardware Details

- Intel STL2 MB ServerWorks Chipset
- 1 Gb Memory
- 1 onbord EIDE controller,
  1 Promise Ultra66 card
- 10 IBM 75Gb , 21 WD 100Gb disks
- 4-port Fastethernet cards
  (D-Link DFE-570TX)
- 24port Fastethernet switch
  (3COM SuperStacker 3300)
- KVM switches, LCD display
- 2U rackmounted enclosures
  (space for 4 disks, 1 floppy, 1 CD)
- Two 12U racks

The System

Installation

overall versions

- FRHL 6.12 kernel 2.2.19
- MOSIX version 0.98.0

Special Gateway features:

- Processes migrate away (job arrival blocked, apparent speed reduced)
- NIS from MIT cluster
- /home nfs mounted
- Cdfsoft installed
- Ethernet channel bonding (3 channels)

Plain Worker Nodes:

- No accounts, no software

Operation

- System in "Production" for over 1.5 years
- MC production
- CDF data samples resident on disk
- Up to 2 Tb of data
- Data analysis by 5-15 active users
- Code development on gateways

Known Problems & Concerns

- Intrinsic problems
  - Home node concept (system call execution)
  - Lost connections
  - Missing special file support
- MOSIX bugs
  - Open file problem
  - Login problem (freeze)
- Maintenance issues
  - MOSIX release/bugfix policy
  - Missing tools (example: no df)

Latest Developments

- New 5 node dual XEON test cluster
- FRHL 7.3 and MOSIX 1.8.0
- System more stable
- Several bugs fixed
- Home node concept still the main limiting factor

Future Plans & Conclusion

- Centralized network boot
- Improve maintenance issues
- Investigate Data File Catalog to organize access
- Watch closely new R&D efforts (SSI, openMosix)

Conclusion:

- Good experience with a small cluster
- System currently not fully suited for a large scale system (judging for version 0.98.0)

Acknowledgements

T.Kim, D.Waters, F.Würthwein

- Benchmarks
- Problem investigation
- New R&D

S.Lammel, H.Matsunaga, MOSIX team, M.Neubauer, T.Vaiciulis

- Useful discussions

S.Tether

- Current maintenance and upgrade