

The Cray XT Programming Environment

Roberto Ansaloni
roberto@cray.com

December 2007

Agenda

- Programming Environment
 - Job launch
 - modules
- Compilers
 - PGI compilers: common flags, optimization
 - Pathscale compilers: common flags, optimization
 - GNU compilers
- I/O optimization
 - I/O architecture overview
 - Lustre features
 - Ifs command
- Running and watching an application
 - ALPS aprun
 - PBSPro

Cray XT programming environment is SIMPLE

- Edit and compile MPI program (no need to specify include files or libraries)

```
$ vi pippo.f  
$ ftn -o pippo pippo.f
```

- Edit PBSPro job file (pippo.job)

```
#PBS -N myjob  
#PBS -l mppwidth=256  
#PBS -j oe  
cd $PBS_O_WORKDIR  
aprun -n 256 ./pippo
```

- Run the job (output will be myjob.oxxxxx)

```
$ qsub pippo.job
```

Cray XT programming environment overview

- PGI compiler suite
- Pathscale compiler suite
- Optimized libraries:
 - 64 bit AMD Core Math library (ACML)
Level 1,2,3 of BLAS, LAPACK, FFT
 - SciLib: Scalapack, BLACS, SuperLU
- MPI-2 message passing library for communication between nodes
(derived from MPICH-2, implements MPI-2 standard, except for support of dynamic process functions)
- SHMEM one-sided communication library

Cray XT programming environment overview

- GNU C library, gcc, g++
- aprun command to launch applications; similar to mpirun command
- PBSPro batch system
- Performance tools: CrayPat, Apprentice2
- Totalview debugger

The module tool on the Cray XT

- How can we get appropriate Compiler and Libraries to work with?
- module tool used on the Cray XT to handle different versions of packages (compiler, tools,...):
 - e.g.: **module load compiler1**
 - e.g.: **module switch compiler1 compiler2**
 - e.g.: **module load totalview**
 -
- taking care of changing of PATH, MANPATH, LM_LICENSE_FILE,.... environment.
- user should not set those environment variable in his shell startup files, makefiles,...
- keep things flexible to other package versions

Cray XT programming environment: module list

```
shark> module list
```

```
Currently Loaded Modulefiles:
```

```
1) modules/3.1.6          7) xt-mpt/2.0.36        13) xt-catamount/2.0.36
2) MySQL/4.0.27          8) xt-pe/2.0.36         14) xt-boot/2.0.36
3) pgi/7.1.3             9) PrgEnv-pgi/2.0.36    15) xt-lustre-ss/2.0.36
4) totalview-support/1.0.2 10) xt-service/2.0.36   16) xtpe-target-cn1
5) xt-totalview/8.3      11) xt-libc/2.0.36      17) Base-opts/2.0.36
6) xt-libsci/10.2.0      12) xt-os/2.0.36       18) pbs/default
```

- Current versions
 - Unicos/lc (CNL) 2.0
 - PGI 7.1
 - ACML 4.0
 - PBS 8.1

Cray XT programming environment: module show

```
shark> module show pgi
```

```
-----  
/opt/modulefiles/pgi/7.1.3:
```

```
setenv          PGI_VERSION 7.1  
setenv          PGI_VERS_STR 7.1.3  
setenv          PGI_PATH /opt/pgi/7.1.3  
setenv          PGI /opt/pgi/7.1.3  
prepend-path    LM_LICENSE_FILE /opt/pgi/7.1.3/license.dat  
prepend-path    PATH /opt/pgi/7.1.3/linux86-64/7.1/bin  
prepend-path    MANPATH /opt/pgi/7.1.3/linux86-64/7.1/man  
prepend-path    LD_LIBRARY_PATH /opt/pgi/7.1.3/linux86-64/7.1/lib  
prepend-path    LD_LIBRARY_PATH /opt/pgi/7.1.3/linux86-64/7.1/libso  
-----
```

Useful module commands

- List available modules
module avail
- Use profiling
module load craypat
- Change PGI compiler version
module swap pgi/7.1.3 pgi/7.0.7
- Load Pathscale environment
module swap PrgEnv-pgi PrgEnv-pathscales

Compiler drivers to create CNL executables

- When the PrgEnv is loaded the compiler drivers are also loaded
 - By default PGI compiler under compiler drivers
 - the compiler drivers also take care of loading appropriate libraries (-lmpich, -lsci, -lacml, -lpapi)

- Available drivers (also for linking of MPI applications):

Fortran 90/95 programs	ftn
Fortran 77 programs	f77
C programs	cc
C++ programs	CC

- Cross compiling environment
 - Compiling on a Linux service node
 - Generating an executable for a CNL compute node
 - Do not use pgf90, pgcc unless you want a Linux executable
 - Information message:

```
ftn: INFO: linux target is being used
```

PGI compiler flags for a first start

Overall Options:

- Mlist creates a listing file
- Mneginfo information on why certain optimizations are not performed.
- WI,-M generates a loader map (to stdout)

Preprocessor Options:

- Mpreprocess runs the preprocessor on Fortran files (default on .F, .F90, or .fpp files)

Optimisation Options:

- fast chooses generally optimal flags for the target platform
- fastsse chooses generally optimal flags for a processor that supports the SSE, SSE3 instructions.
- Mipa=fast,inline Inter Procedural Analysis
- Minline=levels:number number of levels of inlining

man pgf90, man pgcc, man pgCC

PGI User's Guide (Chapter 2) <http://www.pgroup.com/doc/pgiug.pdf>

Other programming environments

- GNU
 - **module swap PrgEnv-pgi PrgEnv-gnu**
 - Default compiler is gcc/3.2.3
 - gcc/4.1.1 module available

- Pathscale
 - **module swap PrgEnv-pgi PrgEnv-pathscale**
 - Pathscale version is 2.5

Pathscale compiler flags for a first start

Preprocessor Options:

-cpp runs cpp on source files
-ftpp runs the fortran source preprocessor

Optimisation Options:

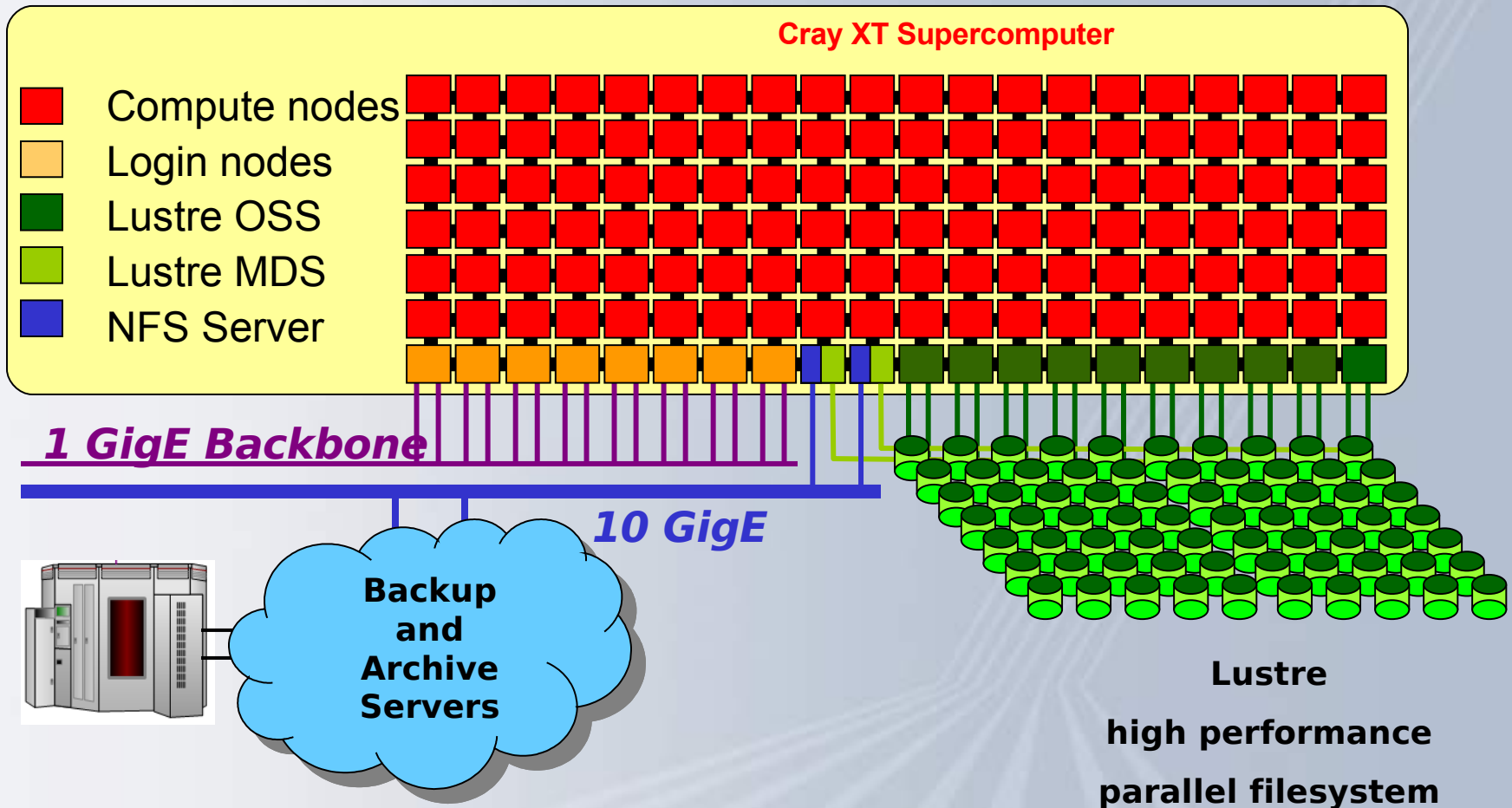
-LNO: specify transformations performed on loop nests by the Loop Nest Optimizer
-OPT: controls miscellaneous optimizations
-ipa Inter Procedural Analysis
-Ofast Equivalent to
-O3 -ipa -OPT:Ofast -fno-math-errno -ffast-math

Default: -O2 -mcpu=opteron -m64 -msse -msse2 -mno-sse3 -mno-3dnow

Start: -O3 -OPT:Ofast

More info: man eko, man pathf95

The Storage Environment

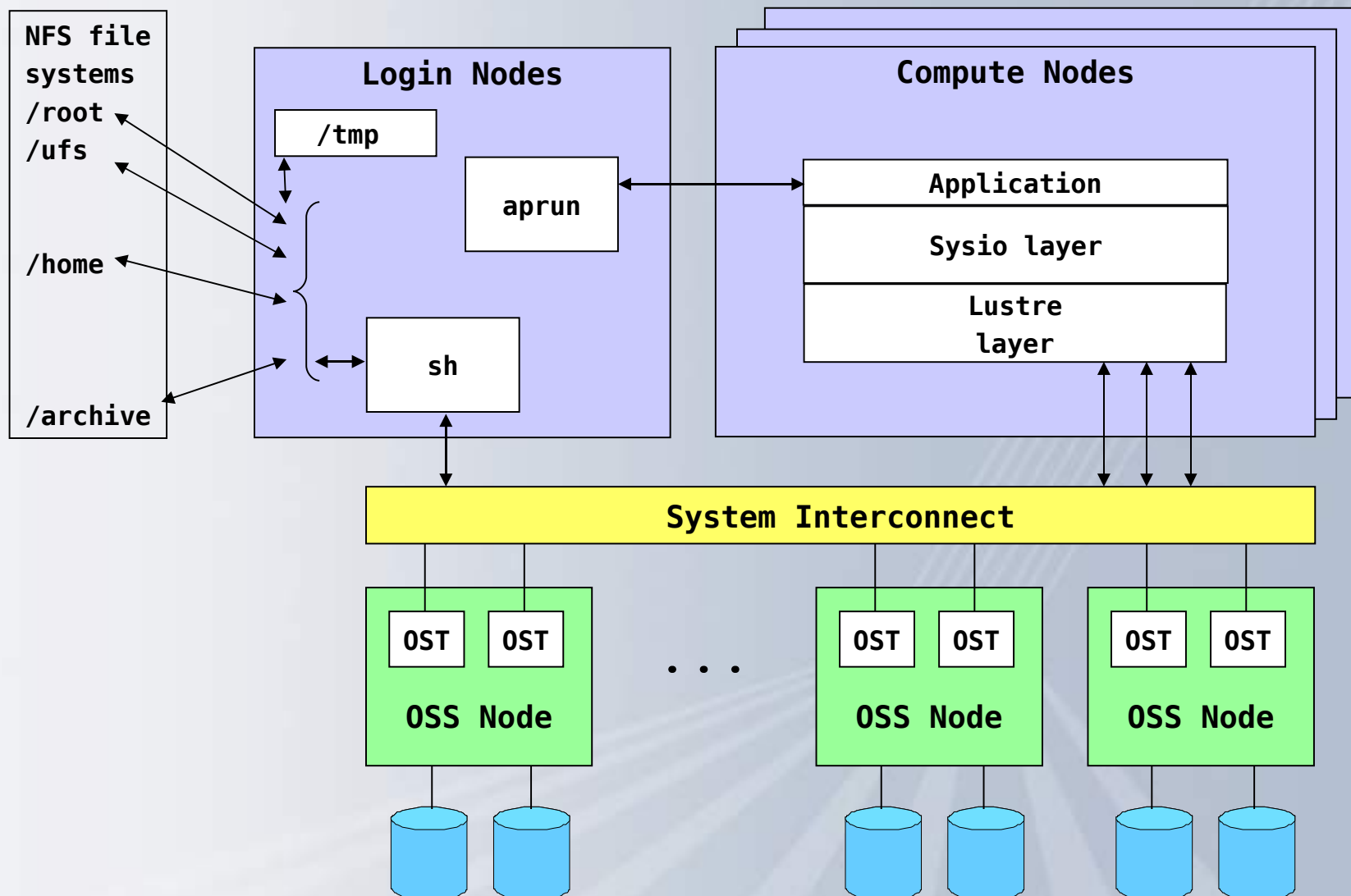


- Cray provides high performance local file system
- Cray enables vendor independent integration for backup and archival

Cray XT I/O architecture

- All I/O is offloaded to service nodes
- Lustre – High performance parallel I/O file system
 - Direct data transfer between compute nodes and files
- aprun and parallel applications can access ONLY lustre
 - The binary file must be on a lustre file system
 - This restrictions will disappear in future releases (DVS I/O forwarding)
- No local disks
- /tmp is a MEMORY file system, on each login node

Cray XT I/O architecture



Lustre



- A scalable cluster file system for Linux
 - Developed by Cluster File Systems, Inc.
 - Name derives from “Linux Cluster”
 - The Lustre file system consists of software subsystems, storage, and an associated network
- Terminology
 - **MDS** – metadata server
 - Handles information about files and directories
 - **OSS** – Object Storage Server
 - The hardware entity
 - The server node
 - Support multiple OSTs
 - **OST** – Object Storage Target
 - The software entity
 - This is the software interface to the backend volume

Lustre File Striping

- Stripes defines the number of OSTs to write the file across
 - Can be set on a per file or directory basis

- CRAY recommends that the default be set to
 - not striping across all OSTs, but
 - set default stripe count of one to four

- But not always the best for application performance.
As a general rule of thumbs :
 - If you have one large file
=> stripe over all OSTs
 - If you have a large number of files (~2 times #OSTs)
=> turn off striping (#stripes=1)

Lustre lfs command

- **lfs** is a lustre utility that can be used to create a file with a specific striping pattern, displays file striping patterns, and find file locations
- The most used options are :
 - setstripe
 - getstripe
 - df
- For help execute **lfs** without any arguments

```
$ lfs
lfs > help
Available commands are:
    setstripe
    find
    getstripe
    check
    ...
```

lfs setstripe

- Sets the stripe for a file or a directory

- **lfs setstripe <file|dir> <size> <start> <count>**
 - stripe size: Number of bytes on each OST (0 filesystem default)
 - stripe start: OST index of first stripe (-1 filesystem default)
 - stripe count: Number of OSTs to stripe over (0 default, -1 all)

- Comments
 - The stripes of a file is given when the file is created. It is not possible to change it afterwards.
 - If needed, use lfs to create an empty file with the stripes you want (like the touch command)

Lustre striping hints

- For maximum aggregate performance: **Keep all OSTs occupied**
- Many clients, many files: **Don't stripe**
If number of clients and/or number of files \gg number of OSTs:
Better to put each object (file) on only a **single** OST.
- Many clients, one file: **Do stripe**
When multiple processes are all accessing one large file:
Better to stripe that single file over **all** of the available OSTs.
- Some clients, few large files: **Do stripe**
When a few processes access large files in large chunks:
Stripe over **enough** OSTs to keep the OSTs busy on both write and read paths.

Running an application on the Cray XT

- ALPS
 - Application Level Placement Scheduler

- aprun is the ALPS application launcher
 - It must be used to run application on the XT compute nodes
 - If aprun is not used, the application is launched on the login node (and likely fails)

Running an application on the Cray XT - aprun

- aprun has (at least) 3 important parameters to control:
 - how many nodes are required
 - how many MPI tasks are used
 - how many cores are used on a multicore chip
 - how many OpenMP threads are required
- These aprun parameters corresponds to PBS parameters
 - If an aprun parameter is specified, also the corresponding PBS one must be specified

aprun	PBS	meaning (simplified)
-n	-l mppwidth	total number of MPI tasks
-N	-l mppnppn	number of MPI tasks per node
-d	-l mppdepth	number of OpenMP threads per node

Running an application on the Cray XT - examples

- Assuming a dual-core system (2 cores per node)
- Pure MPI application, using all the available cores in a node
 - npes MPI tasks, npes cores, npes/2 nodes

```
$ aprun -n <npes>
```

- Pure MPI application, using only 1 core per node
 - npes MPI tasks, 2*npes cores allocated, npes nodes allocated
 - you may want to do this to give all the memory on the node to the MPI tasks

```
$ aprun -N 1 -n <npes>
```

- Hybrid MPI/OpenMP application
 - npes MPI tasks, 2 OpenMP threads each, 2*npes cores, npes nodes
 - need to set OMP_NUM_THREADS

```
$ export OMP_NUM_THREADS=2  
$ aprun -N 1 -d 2 -n <npes>
```

Remember to set the corresponding PBS option in the job header

Running an application on the Cray XT - MPMD

- aprun supports MPMD – Multiple Program Multiple Data
 - Launching several executables on the same MPI_COMM_WORLD

```
$ aprun -n 128 exe1 : -n 64 exe2 : -n 64 exe3
```

Running a batch application with PBSPro

- The number of required nodes and cores is determined by the parameters specified in the job header

```
#PBS -l mppwidth=256
```

```
#PBS -l mppnppn=1
```

- The job is submitted by the qsub command
- At the end of the execution output and error files are returned to submission directory
- PBS environment variable: \$PBS_O_WORKDIR
Set to the directory from which the job has been submitted

Other PBSPro parameters

- `#PBS -N job_name`
the job name is used to determine the name of job output and error files
- `#PBS -l walltime=hh:mm:ss`
Maximum job elapsed time
should be indicated whenever possible: this allows PBS to determine best scheduling strategy
- `#PBS -j oe`
job error and output files are merged in a single file
- `#PBS -q queue`
request execution on a specific queue: usually not needed

MPI Dual-core vs MPI Single-core vs Hybrid MPI+OpenMP

All the examples allocate 128 dual core nodes

DUAL CORE (most common)

```
#PBS -N DCjob
#PBS -l mppwidth=256
#PBS -j oe

aprun -n 256 pippo
```

SINGLE CORE

```
#PBS -N SCjob
#PBS -l mppwidth=128
#PBS -l mppnppn=1
#PBS -j oe

aprun -N1 -n 128 pippo
```

Hybrid MPI + OpenMP

```
#PBS -N OMPjob
#PBS -l mppwidth=128
#PBS -l mppnppn=1
#PBS -l mppdepth=2
#PBS -j oe

export OMP_NUM_THREADS=2
aprun -N1 -d2 -n 128 pippo
```

Watching a launched job on the Cray XT

- `xtshowcabs`
 - Shows XT nodes allocation and aprun processes
 - Both interactive and PBS

- `apstat`
 - Shows aprun processes status
 - `apstat` overview
 - `apstat -a [apid]` info about all the applications or a specific one
 - `apstat -n` info about the status of the nodes

- PBS `qstat` command
 - shows batch PBS jobs
 - `qstat -r` check running jobs
 - `qstat -n` check running and queued jobs
 - `qstat -s <job_id>` reports comments about the job

Which processors am I using ?

- xtprocadmin
- xtshowcabs
- ALPS allocation strategy
- XT flat performance machine

xtprocadmin

```
perch> xtprocadmin -A
```

NID	(HEX)	NODENAME	TYPE	ARCH	OS	CORES	AVAILMEM	PAGESZ	CLOCKMHZ
0	0x0	c0-0c0s0n0	service	xt	(service)	1	2000	4096	2400
3	0x3	c0-0c0s0n3	service	xt	(service)	1	2000	4096	2400
4	0x4	c0-0c0s1n0	service	xt	(service)	1	2000	4096	2400
7	0x7	c0-0c0s1n3	service	xt	(service)	1	2000	4096	2400
8	0x8	c0-0c0s2n0	service	xt	(service)	1	2000	4096	2400
11	0xb	c0-0c0s2n3	service	xt	(service)	1	2000	4096	2400
12	0xc	c0-0c0s3n0	service	xt	(service)	1	2000	4096	2400
15	0xf	c0-0c0s3n3	service	xt	(service)	1	2000	4096	2400
16	0x10	c0-0c0s4n0	compute	xt	CNL	1	2000	4096	2400
17	0x11	c0-0c0s4n1	compute	xt	CNL	1	2000	4096	2400
18	0x12	c0-0c0s4n2	compute	xt	CNL	1	2000	4096	2400
19	0x13	c0-0c0s4n3	compute	xt	CNL	1	2000	4096	2400
20	0x14	c0-0c0s5n0	compute	xt	CNL	1	2000	4096	2400

xtshowcabs

	C0-0	C1-0	C2-0	C3-0	C4-0	C5-0	C6-0	C7-0
n3	iiiXiiii	onqqoggg		wwwyyyyy	nnnBBBBB	DDDDDDxB	CCCzzzzz	GGGGGGgw
n2	iiiiiii	inqqoggg		wwwyyyyy	nnnzBBBB	DDDDxB	CCCCzzzz	GGGGGGgw
n1	iiiiiii	nnnrqogg		wwwyyyyy	nnnnBBBB	DDDDxB	CCCCzzzz	GGGGGGgw
c2n0	iiiiiii	npqqoggg		wwwxyyyy	nnnnBBBB	DDDDx	BCCCzzzz	wGGGGGGX
n3	aeggiii	iiiiinn	ggsitv	q ppw	nnnnnnnn		yyyBBBBB	zzzzzBwF
n2	adggiii	kiiiiinn	ggsout	qw npw	nnnnnnnn		yyyBBBBB	zzzzzBwF
n1	acggiii	jiiilinn	ggsott	qq w	nnnnnnnn		yyyBBBBB	zzzzzBwF
c1n0	abfgghii	iiiiinn	ggsott	qq q	nnnnnnnn		yyyBBBBB	zzzzzzwF
n3	SSSSSSS:	SSSSSS::	gggggggg	qqq	yyyyyyyn	qpppCBB	BBBwwwy	xEzzzzzz
n2	:	::	gggggggg	qqq	yyyyyyyn	qqpppBB	BBBwwwy	xEzzzzzz
n1	:	::	gggggggg	qq	yyyyyyyn	qqpppBB	BBBwww	xEzzzzzz
c0n0	SSSSSSS:	SSSSSS::	gggggggg	qq	yyyyyyyn	qqpppCBB	BBBwww	zEzzzzzz
s	01234567	01234567	01234567	01234567	01234567	01234567	01234567	01234567

Cabinet 3

nodename: c3

xtshowcabs

	C0-0	C1-0	C2-0	C3-0	C4-0	C5-0	C6-0	C7-0
n3	iiiXiiii	onqqoggg		wwwyyyyy	nnnBBBBB	DDDDDDxB	CCCzzzzz	GGGGGGgw
n2	iiiiiii	inqqoggg		wwwyyyyy	nnnzBBBB	DDDDxB	CCCCzzzz	GGGGGGgw
n1	iiiiiii	nnnrqogg		wwwyyyyy	nnnnBBBB	DDDDxB	CCCCzzzz	GGGGGGgw
c2n0	iiiiiii	npqqoggg		wwwxyyyy	nnnnBBBB	DDDDx	BCCCzzzz	wGGGGGGX
n3	aeggiii	iiiiinn	ggsitv	q ppw	nnnnnnnn		yyyBBBBB	zzzzzBwF
n2	adggiii	kiiiiinn	ggsout	qw npw	nnnnnnnn		yyyBBBBB	zzzzzBwF
n1	acggiii	jiiilinn	ggsott	qq w	nnnnnnnn		yyyBBBBB	zzzzzBwF
c1n0	abfgghii	iiiiinn	ggsott	qq q	nnnnnnnn		yyyBBBBB	zzzzzzwF
n3	SSSSSSS:	SSSSSS::	gggggggg	qqq	yyyyyyyn	qpppCBB	BBBwwwwy	xEzzzzzz
n2	:	::	gggggggg	qqq	yyyyyyyn	qqpppBB	BBBwwwwy	xEzzzzzz
n1	:	::	gggggggg	qq	yyyyyyyn	qqpppBB	BBBwwwwy	xEzzzzzz
c0n0	SSSSSSS:	SSSSSS::	gggggggg	qq	yyyyyyyn	qqpppCBB	BBBwwwwy	zEzzzzzz
s	01234567	01234567	01234567	01234567	01234567	01234567	01234567	01234567

Cabinet 3, chassis 1

nodename: c3-0c1

xtshowcabs

	C0-0	C1-0	C2-0	C3-0	C4-0	C5-0	C6-0	C7-0
n3	iiiXiiii	onqqoggg		wwwyyyyy	nnnBBBBB	DDDDDDxB	CCCzzzzz	GGGGGGgw
n2	iiiiiii	inqqoggg		wwwyyyyy	nnnzBBBB	DDDDxB	CCCCzzzz	GGGGGGgw
n1	iiiiiii	nnnrqogg		wwwyyyyy	nnnnBBBB	DDDDxB	CCCCzzzz	GGGGGGgw
c2n0	iiiiiii	npqqoggg		wwwxyyyy	nnnnBBBB	DDDDxD	BCCCzzzz	wGGGGGGX
n3	aeggiii	iiiiinn	ggsitv	q pw	nnnnnnnn		yyyBBBBB	zzzzzBfw
n2	adggiii	kiiiiinn	ggsout	qw nw	nnnnnnnn		yyyBBBBB	zzzzzBwF
n1	acggiii	jiiilinn	ggsott	qq w	nnnnnnnn		yyyBBBBB	zzzzzBwF
c1n0	abfgghii	iiiiinn	ggsott	qq g	nnnnnnnn		yyyBBBBB	zzzzzzwF
n3	SSSSSSS:	SSSSSS::	gggggggg	qqq	yyyyyyyn	qpppCBB	BBBwwwwy	xEzzzzzz
n2	:	::	gggggggg	qqq	yyyyyyyn	qqpppBB	BBBwwwwy	xEzzzzzz
n1	:	::	gggggggg	qq	yyyyyyyn	qqpppBB	BBBwwwwy	xEzzzzzz
c0n0	SSSSSSS:	SSSSSS::	gggggggg	qq	yyyyyyyn	qqpppCBB	BBBwwwwy	zEzzzzzz
s	01234567	01234567	01234567	01234567	01234567	01234567	01234567	01234567

Cabinet 3, chassis 1, slot 6

nodename: c3-0c1s6

xtshowcabs

	C0-0	C1-0	C2-0	C3-0	C4-0	C5-0	C6-0	C7-0
n3	iiiXiiii	onqqoggg		wwwyyyyy	nnnBBBBB	DDDDDDxB	CCCzzzzz	GGGGGGgw
n2	iiiiiii	inqqoggg		wwwyyyyy	nnnzBBBB	DDDDxB	CCCCzzzz	GGGGGGgw
n1	iiiiiii	nnnrqogg		wwwyyyyy	nnnnBBBB	DDDDxB	CCCCzzzz	GGGGGGgw
c2n0	iiiiiii	npqqoggg		wwwxyyyy	nnnnBBBB	DDDDx	BCCCzzzz	wGGGGGGX
n3	aeggiii	iiiiinn	ggsitv	q p	nnnnnnnn		yyyBBBBB	zzzzzBfW
n2	adggiii	kiiiiinn	ggsout	qw n	nnnnnnnn		yyyBBBBB	zzzzzBwF
n1	acggiii	jiiilinn	ggsott	qq v	nnnnnnnn		yyyBBBBB	zzzzzBwF
c1n0	abfgghii	iiiiinn	ggsott	qq g	nnnnnnnn		yyyBBBBB	zzzzzzwF
n3	SSSSSSS:	SSSSSS::	gggggggg	qqq	yyyyyyyn	qpppCBB	BBBwwwy	xEzzzzzz
n2	:	::	gggggggg	qqq	yyyyyyyn	qqpppBB	BBBwwwy	xEzzzzzz
n1	:	::	gggggggg	qq	yyyyyyyn	qqpppBB	BBBwww	xEzzzzzz
c0n0	SSSSSSS:	SSSSSS::	gggggggg	qq	yyyyyyyn	qqpppCBB	BBBwww	zEzzzzzz
s	01234567	01234567	01234567	01234567	01234567	01234567	01234567	01234567

Cabinet 3, chassis 1, slot 6, node 2

nodename: c3-0c1s6n2

nid: 442 (0x1ba)

xtshowcabs: service nodes

```

      C0-0      C0-1      C1-0      C1-1      C2-0      C2-1      C3-0      C3-1
n3 bbbbeeee aaccccc ihhiihc bbbbjjb ccccccc ooooo111 ddnnlenn dnnnn111
n2 bbbbeeee aaccccc ihhiihc bbbbjbj ccccccc ooooo11d ddnnlknn dnnnn111
n1 bbbbeeee aaccccc gihhiih bbbbjbj ccccccc ooooo111 dddnllnn ddnnnn11
c2n0 bbbbeeee aaccccc hiihiih bbbbjbj ccccccc ooooo011 dddnnlenn ddnnnn11
n3 bdddbcc gggggga gggghhh gggfgfb cddddd bbooooo dddddd nggpidd
n2 bdddbcc gggggga gggghhh gggggfb cddddd bbooooo dddddd ndgphida
n1 bdddbcc gggggga gggghhh gggggfj cddddd bbooooo dddddd nngghidn
c1n0 bdddbcc gggggga gggghhh gggggfj cddddd bbooooo dddddd nngghidd
n3 SSSSSSb eeefbfg SSSSSScc ccccccg jmmmbmbb ccnnnnd dddddd Snnnnn
n2 SSSSSSb eeefbfg SSSSSScc cccfcg jmbmbb ccnnnnd dddddd nnnnn
n1 SSSSSSb eeefbfg SSSSSScc ccccccc jlmmbmbb ccnnnnd dddddd nnnnn
c0n0 SSSSSSa eeefbfg SSSSSScc ccccccc jkmmmmmm ccnnnnd dddddd Snnnnn
s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567

```

Legend:

- | | |
|---------------------------------|------------------------------------|
| nonexistent node | S service node |
| : free interactive compute node | A allocated, but idle compute node |
| free batch compute node | ? suspect compute node |
| X down compute node | Y down or admin down service node |
| Z admin down compute node | R node is routing |

xtshowcabs: free batch nodes

```

      C4-0      C4-1      C5-0      C5-1      C6-0      C6-1      C7-0      C7-1
n3 lppppppp pppppppp ppnnllll iiiiiiiii llllllqq ssssssss bbuvvvvw BBB| |||
n2 lppppppp pppppppp ppnnllll iiiiiiiii llllllqq ssssssss bbuuvvvv BBBB |||
n1 llpppppp pppppppp ppinllll iiiiiiiii llllllqq ssssssss ubuuvvvv BBBB |||
c2n0 llpppppp pppppppp ppinnlll iiiiiiiii llllllqq ssssssss bbbuvvvv BBBB |||
n3 laalllll pppppppp pppppppp iiiiiiiii ggnnnnll ggssssss tttuguub yyyzzzzB
n2 llalllll pppppppp pppppppp iiiiiiiii ggnnnnll ggssssss ttttguub yyyzzzz
n1 llalllll pppppppp pppppppp iiiiiiiii ggnnnnll ggssssss ttttguub yyyzzzz
c1n0 llaallll pppppppp pppppppp iiiiiiiii gglnnnnl ggssssss ttttgguu yyyzzzz
n3 llllllaa Sppppppp pppppppp nniiiiii iiiigggg qqrrgggg sssskkkk wwwxxxxy
n2 llllllaa ppppppp pppppppp nniiiiii iiiigggg qqrrgggg sssskkk| wwwxxxxx
n1 llllllaa ppppppp pppppppp nniiiiii iiiigggg qqgrggrg sssskkkk wwwxxxxx
c0n0 llllllaa Sppppppp pppppppp nniiiiii iiiigggg qqgrrggg sssskkkk wwwxxxxx
      s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567

```

Legend:

- nonexistent node
- S service node
- : free interactive compute node
- A allocated, but idle compute node
- | free batch compute node
- ? suspect compute node
- X down compute node
- Y down or admin down service node
- Z admin down compute node
- R node is routing

xtshowcabs: down compute nodes

```

C0-0   C1-0   C2-0   C3-0   C4-0   C5-0   C6-0   C7-0
n3 aaaaaaaa dddddd ggghhhh hhhhhiii hhhhhhhh iiiiii iiiiii iiiiii
n2 aaaaaaaa dddddd ggghhhh hhhhhiii hhhhhhhh iiiiii iiiiii iiiiii
n1 aaaaaaaa dddddd ggghhhh hhhhhiii hhhhhhhh iiiiii iiiiii iiiiii
c2n0 aaaaaaaa dddddd ggghhhh hhhhhiii hhhhhhhh iiiiii iiiiii iiiiii
n3 ::aaaaba aaaacccc ffffffff hhhhhhhh hhhhhhhh hhhhhhi iiiiii iiiiii
n2 ::aaaaba aaaacccc ffffffff hhhhhhhh hhhhhhhh hhhhhhi iiiiii iiiiii
n1 ::aaaaba aaaacccc ffffffff hhhhhhhh hhhhhhhh hhhhhhi iiiiii iiiiii
c1n0 ::aaaaba aaaacccc ffffffff hhhhhhhh hhhhhhhh hhhhhhi iiiiii iiiiii
n3 SSSSSS:: SSSSaaa deeeeeeef hhhhhhhh iihhhhh hhhhhhhh iiiiii iiiiii
n2      ::      aaa deeeeeeef hhhhhhhh iihhhhh hhhhhhhh iiiiii iiiiii
n1      ::      aaa deeeeeeef hhhhhhhh iihhhhh hhhhhhhh iiiiii iiiiii
c0n0 SSSSSS:: SSSSaaa deeeeeeef hhhhhhhh iihhhhh hhhhhhhh iiiiii iiiiii
s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567

```

Sorry, could not find any of them

```

C8-0   C9-0   C10-0
n3 jjjjjjjf kkk| | | | | | | |
n2 jjjjjjjf kkk| | | | | | | |
n1 jjjjjjjf kkk| | | | | | | |
c2n0 jjjjjjjf kkk| | | | | | | |
n3 jjjjjjjj | | | | | | | | k | | | | | | | |
n2 jjjjjjjj | | | | | | | | k | | | | | | | |
n1 jjjjjjjj | | | | | | | | k | | | | | | | |
c1n0 jjjjjjjj | | | | | | | | k | | | | | | | |
n3 hhhggggj ffffffff| | | | | | | |
n2 hhhggggj ffffffff| | | | | | | |
n1 hhhggggj ffffffff| | | | | | | |
c0n0 hhhggggj ffffffff| | | | | | | |
s01234567 01234567 01234567

```

Legend:

X down compute node

Y down or admindown service node

Z admindown compute node

R node is routing

ALPS allocation algorithm

- The first available compute processors are selected, scanning the processor list sequentially by NID
- NID sequence has no relationship with the XT topology

```
$ xtprocadmin | grep compute | grep batch | grep up | grep '4$' | head -10
```

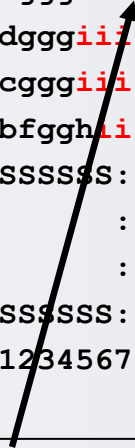
206	0xce	c1-0c2s3n2	compute	up	batch	4	4
207	0xcf	c1-0c2s3n3	compute	up	batch	4	4
208	0xd0	c1-0c2s4n0	compute	up	batch	4	4
209	0xd1	c1-0c2s4n1	compute	up	batch	4	4
210	0xd2	c1-0c2s4n2	compute	up	batch	4	4
211	0xd3	c1-0c2s4n3	compute	up	batch	4	4
212	0xd4	c1-0c2s5n0	compute	up	batch	4	4
213	0xd5	c1-0c2s5n1	compute	up	batch	4	4
214	0xd6	c1-0c2s5n2	compute	up	batch	4	4
215	0xd7	c1-0c2s5n3	compute	up	batch	4	4

Processor allocation to applications

```

C0-0    C1-0    C2-0    C3-0    C4-0    C5-0    C6-0    C7-0
n3 iiiXiiii onqqoggg ||||||| wwwyyyyy nnnBBBBB DDDDDx B CCCzzzzz GGGGGGww
n2 iiiiiiii inqqoggg ||||||| wwwyyyyy nnnzBBBB |DDDDxB CCCCzzzz GGGGGGww
n1 iiiiiiii nnnrqogg ||||||| wwwyyyyy nnnnBBBB |DDDDxB CCCCzzzz GGGGGGww
c2n0 iiiiiiii npqqoggg ||||||| wwwxyyyy nnnnBBBB |DDDDxD BCCCzzzz wGGGGGGX
n3 aegggiii iiiiinn ggsitv| | q|||ppw nnnnnnnn ||||||| yyyBBBBB zzzzzBwF
n2 adgggiii kiiiiinn ggsout| | qw|||npw nnnnnnnn ||||||| yyyBBBBB zzzzzBwF
n1 acgggiii jiiilinn ggsott| | qq|||w nnnnnnnn ||||||| yyyBBBBB zzzzzBwF
c1n0 abfgghi iiiiinn ggsott| | qq|||q nnnnnnnn ||||||| yyyBBBBB zzzzzzwF
n3 SSSSSSS: SSSSSS:: gggggggg |||||qqq yyyyyyyyn qpppCBB| BBBwwwwy xEzzzzzz
n2      :      :: gggggggg |||||qqq yyyyyyyyn qpppBB| BBBwwwwy xEzzzzzz
n1      :      :: gggggggg |||||qqq yyyyyyyyn qpppBB| BBBBwww xEzzzzzz
c0n0 SSSSSSS: SSSSSS:: gggggggg |||||qqq yyyyyyyy qpppCBB BBBBwww zEzzzzzz
s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567

```



YODS LAUNCHED ON CATAMOUNT NODES

Job ID	User	Size	Start	yod command line and arguments
i 70609	ymantz	64	Feb 8 14:03:07	yod -size 64 ../RUN/cp2k.popt

X dimension links

```

      C0-0      C1-0      C2-0      C3-0      C4-0      C5-0      C6-0      C7-0
n3 iiiXiiii onqqoggg ||||||| wwwyyyyy nnnBBBBB DDDDDDxB CCCzzzzz GGGGGGww
n2 iiiiiiiii inqqoggg ||||||| wwwyyyyy nnnzBBBB |DDDDDxB CCCCzzzz GGGGGGww
n1 iiiiiiii nnnrqggg ||||||| wwwyyyyy nnnnBBBB |DDDDB CCCCzzzz GGGGGCww
-----
c2n0 iiiiiiiii npqqoggg ||||||| wwxyyyyy nnnnBBBB |DDDDDDx BCCCzzzz wGGGGGGX
  n3 aegggiii iiiiiinn ggsitv| | q|||ppw nnnnnnnn ||||||| yyyBBBBB zzzzzBwF
  n2 adgggiii kiiiiinn ggsout| | qw|||npw nnnnnnnn ||||||| yyyBBBBB zzzzzBwF
  n1 acgggiii jiiilinn ggsott| | qq|||w nnnnnnnn ||||||| yyyBBBBB zzzzzBwF
c1n0 abfgghii iiiiiinn ggsott| | qq|||q nnnnnnnn ||||||| yyyBBBBB zzzzzzwF
  n3 SSSSSSS: SSSSSS:: gggggggg |||||qqq yyyyyyyyn qpppCBB| BBBwwwy xEzzzzzz
  n2      :      :: gggggggg |||||qqq yyyyyyyyn qpppBB| BBBwwwy xEzzzzzz
  n1      :      :: gggggggg |||||qqq yyyyyyyyn qpppBB| BBBBwww xEzzzzzz
c0n0 SSSSSSS: SSSSSS:: gggggggg |||||qqq yyyyyyyyn qpppCBB BBBBwww zEzzzzzz
      s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567

```

Legend:

- | | |
|---------------------------------|------------------------------------|
| nonexistent node | S service node |
| : free interactive compute node | A allocated, but idle compute node |
| free batch compute node | ? suspect compute node |
| X down compute node | Y down or admin down service node |
| Z admin down compute node | R node is routing |

N.B. This really depends on the system topology class, i.e. on the size of the system

Y dimension links

```

      C0-0      C1-0      C2-0      C3-0      C4-0      C5-0      C6-0      C7-0
n3  iiiXiiii  onqqoggg  ||||||||  wwwyyyY  nnnBBBBB  DDDDDx  CCCzzzzz  GGGGGGww
n2  iiiiiiiii  inqqoggg  ||||||||  wwwyyyY  nnnzBBBB  |DDDDxB  CCCCzzzz  GGGGGGww
n1  iiiiiiiii  nnnrqogg  ||||||||  wwwyyyY  nnnnBBBB  |DDDDxB  CCCCzzzz  GGGGGGww
c2n0 iiiiiiiii  npqqoggg  ||||||||  wwwxyyY  nnnnBBBB  |DDDDx  BCCCzzzz  wGGGGGGX
n3  aeggiii  iiiiin  ggsitv|  q||||p  nnnnnnnn  ||||||  yyyBBBBB  zzzzzBwF
n2  adgggiii  kiiimn  ggsout|  qw||||  nnnnnnnn  ||||||  yyyBBBBB  zzzzzBwF
n1  acgggiii  jiiilin  ggsott|  qq||||  nnnnnnnn  ||||||  yyyBBBB  zzzzzBwF
c1n0 abfgghii  iiiiin  ggsott|  qq||||  nnnnnnnn  ||||||  yyyBBBB  zzzzzzwF
n3  SSSSSSS:  SSSSSS:  gggggggg  |||||q  YYYYYYYn  qpppCBB  BBBwww  xEzzzzzz
n2  :          :  gggggggg  |||||q  YYYYYYYn  qpppBB  BBBwww  xEzzzzzz
n1  :          :  gggggggg  |||||q  YYYYYYYn  qpppBB  BBBwww  xEzzzzzz
c0n0 SSSSSSS:  SSSSSS:  gggggggg  |||||q  YYYYYYYn  qpppCBB  BBBwww  zEzzzzzz
      s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567


```

Legend:

- | | |
|---------------------------------|------------------------------------|
| nonexistent node | S service node |
| : free interactive compute node | A allocated, but idle compute node |
| free batch compute node | ? suspect compute node |
| X down compute node | Y down or admin down service node |
| Z admin down compute node | R node is routing |

Z dimension links

```

      C0-0      C1-0      C2-0      C3-0      C4-0      C5-0      C6-0      C7-0
n3  iiiXiiii  onqqoggg  |||||||  wwwyyyyy  nnnBBBBB  DDDDDx  CCCzzzzz  GGGGGGww
n2  iiiiiii  inqqoggg  |||||||   nnnzBBBB  |DDDDxB  CCCCzzzz  GGGGGGww
n1  iiiiiii  nnnrqogg  |||||||  wwwyyyyy  nnnnBBBB  |DDDDxB  CCCCzzzz  GGGGGGww
c2n0 iiiiiii  npqqoggg  |||||||  wwwxyyyy  nnnnBBBB  |DDDDx  BCCCzzzz  wGGGGGGX
n3  aeggiii  iiiiin  ggsitv|  q|||ppw  nnnnnnn  |||||||  yyyBBBBB  zzzzzBFw
n2  adgggiii  kiiimn  ggsout|  qw||npw  nnnnnnn  |||||||  yyyBBBBB  zzzzzBwF
n1  acgggiii  jiiilin  ggsott|  qq|||w  nnnnnnn  |||||||  yyyBBBB  zzzzzBwF
c1n0 abfgghii  iiiiin  ggsott|  qq|||q  nnnnnnn  |||||||  yyyBBBB  zzzzzzwF
n3  SSSSSS:  SSSSSS:  gggggggg  ||||qqq  yyyyyyy  qpppCBB  |BBBwww  xEzzzzz
n2  :  :  gggggggg  ||||qqq  yyyyyyy  qpppBB  |BBBwww  xEzzzzz
n1  :  :  gggggggg  ||||qq  yyyyyyy  qpppBB  |BBBwww  xEzzzzz
c0n0 SSSSSS:  SSSSSS:  gggggggg  ||||qq  yyyyyyy  qpppCBB  BBBBwww  zEzzzzz
      s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567

```

Legend:

nonexistent node	S service node
: free interactive compute node	A allocated, but idle compute node
free batch compute node	? suspect compute node
X down compute node	Y down or admin down service node
Z admin down compute node	R node is routing

Processor allocation to applications

Processor (MPI rank) is not topology correlated

Changes chassis

Starts here

	C0-0	C1-0
n3	482X9371	onqqoggg
n2	37158260	2nqqoggg
n1	26047159	nnnrqogg
c2n0	15936048	npqqoqgg
n3	aeggg260	371581nn
n2	adggg159	k6047mnn
n1	acggg048	j59310nn
c1n0	abfggh37	248269nn
n3	SSSSSSS:	SSSSSSS::
n2	:	::
n1	:	::
c0n0	SSSSSSS:	SSSSSSS::
	s01234567	01234567

Processors allocation does not matter so much

- Nodes allocation strategy is not topology aware
 - Same strategy on every XT system (by NID)
 - Topology depends on the size (class)

- However application performance generally does not suffer from that
 - Reproducible results on production workload
 - The Cray XT provides flat performance

- If you really need it, a specific node list may be requested
 - Both from aprun or PBS
 - aprun: -L
 - PBS: -l mppnodes

Online Cray docs

<http://docs.cray.com/>

http://docs.cray.com/cgi-bin/craydoc.cgi?mode=SiteMap;f=xt3_sitemap