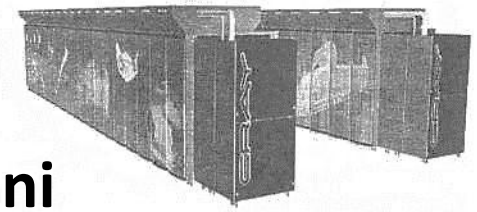
# The Cray Gemini Interconnect



CRAY THE SUPERCOMPUTER COMPANY

H L R S

## Эволюция сети Cray

**SeaStar (Cray XT)**

- Built for scalability to 250K+ cores
- Very effective routing and low contention switch

**Gemini (Cray XE)**

- 100x improvement in message throughput
- 3x improvement in latency
- PGAS Support, Global Address Space
- Scalability to 1M+ cores

**Aries**

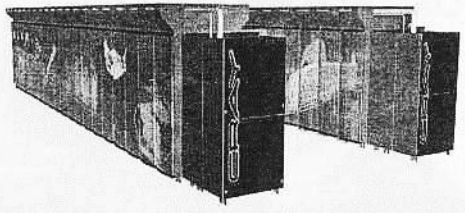
DON'T ask me about it

CRAY THE SUPERCOMPUTER COMPANY

H L R S



# Конфигурация Cray XE6

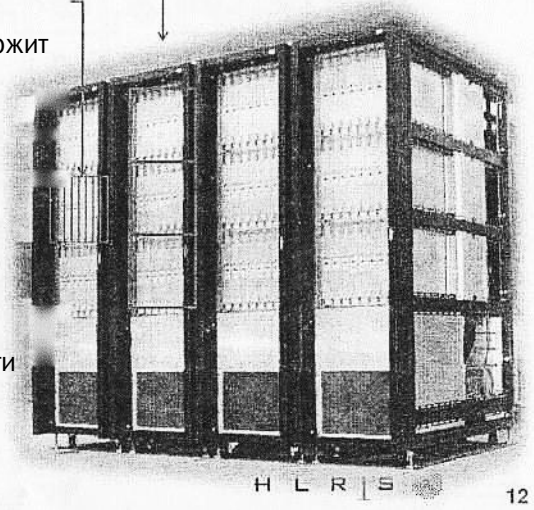


CRAY THE SUPERCOMPUTER COMPANY

H L R S

## Сведения о конфигурации XE6

- XE6 шкаф содержит 3 клетки
- Клетка содержит 8 плат
- Вычислительная плата содержит
  - 8 сокетов
  - 2 соединения Gemini
  - память
  - LO-контроллер
  - VRMs
  - Не движущиеся части
- 1 вентилятор в нижней части

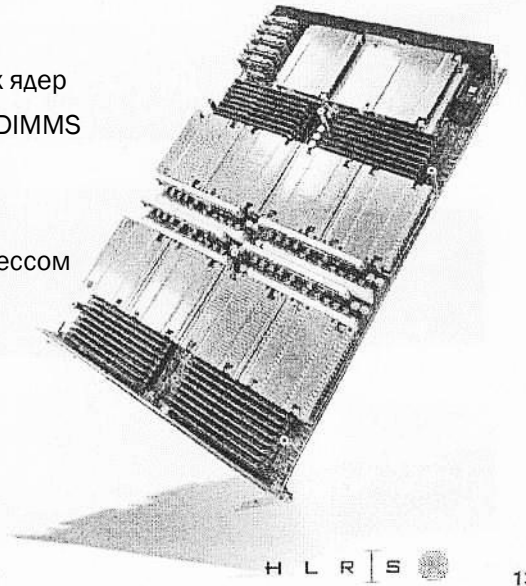


CRAY THE SUPERCOMPUTER COMPANY

H L R S

## Вычислительная плата Cray XE6

- 8 сокетов Magny Cours
- 64 или 96 вычислительных ядер
- 32 модулей памяти DDR3 DIMMS
- 32 канала памяти DDR3
- 2 Gemini ASICs
- LO плата управления процессом



CRAY THE SUPERCOMPUTER COMPANY H L R S



## Топология XE6

**Class 0 Топология {HLRS phase 0)**

Для системы до 3-х секций, с 1 по 9 блоков. Топология представляет собой цельный 3D Торус размером 3N x 4 x 8, где N – число блоков.

**Class 1 Топология**

Для системы от 4-х секций и до 16-ти в одном ряду. Топология представляет собой цельный 3D Торус размером N x 12 x 8, где N – число секций.

**Class 2 Топология (HLRS phase 1)**

Для систем, состоящих из двух рядов. Топология представляет собой цельный 3D Торус размером N x 12 x 16, где N – число секций в ряду (в итоге 2 x N секций). К данному классу отнесены конфигурации из 16 (N = 8) до 48 (N = 24) секций.

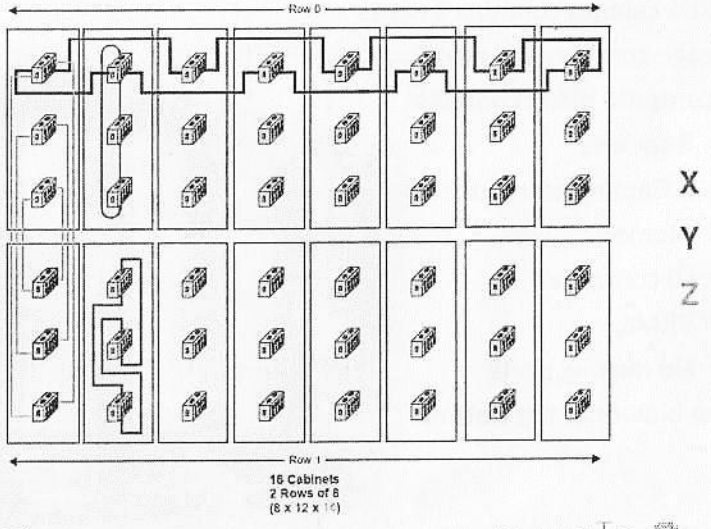
**Class 3 Topology (ORNL JaguarPF)**

Для более крупных, многорядных систем с четным числом рядов. Топология представляет собой цельный 3D Торус размером N (4 x номера рядов) x 24. К данному классу отнесены конфигурации из 48 (4 ряда, 12 секций в ряду) до 576 (12 рядов, 48 секций в ряд) секций.

CRAY H L R S



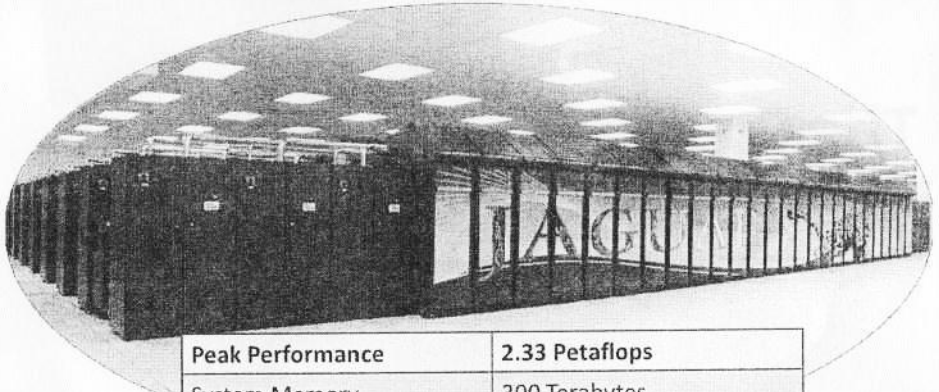
## Class 2 пример : 16 секций, 8 x 12 x 16




CRAY H L R S



## Class 3: ORNL JaguarPF



| | |
|---|---|
| Peak Performance | 2.33 Petaflops |
| System Memory | 300 Terabytes |
| Disk Space | 10.7 Petabytes |
| Interconnect | 3D Torus 25x32x24 |
| Processor Cores | 224,256 |

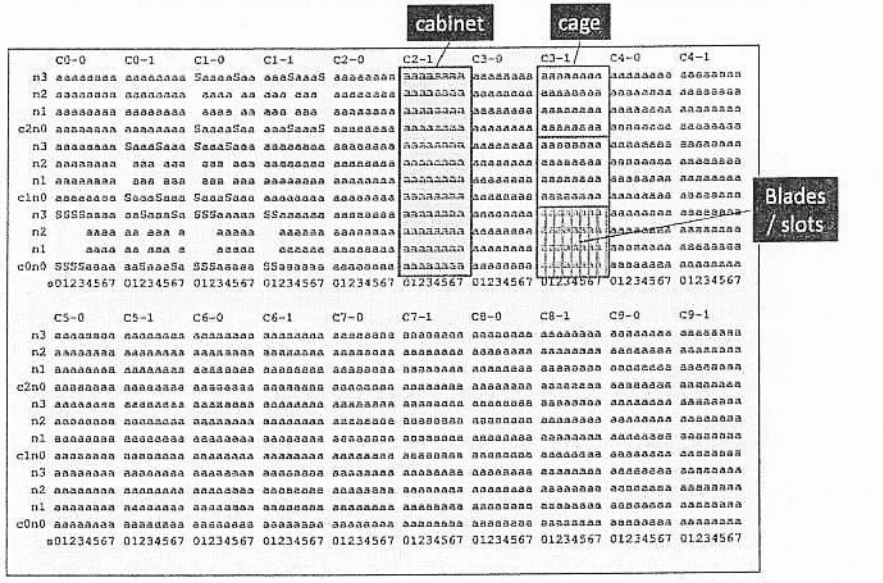
CRAY H L R S

## Описание узлов: xtprocadmin -A

| NID | (HEX) | NODENAME | TYPE | ARCH | OS | CORES | AVAILMEM | PAGESZ | CLOCKMHZ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0x0 | c0-0c0s0n0 | service | xt | (service) | 2 | 8000 | 4096 | 2600 |
| 3 | 0x3 | c0-0c0s0n3 | service | xt | (service) | 2 | 8000 | 4096 | 2600 |
| 4 | 0x4 | c0-0c0s1n0 | service | xt | (service) | 2 | 8000 | 4096 | 2600 |
| 7 | 0x7 | c0-0c0s1n3 | service | xt | (service) | 2 | 8000 | 4096 | 2600 |
| 8 | 0x8 | c0-0c0s2n0 | service | xt | (service) | 2 | 8000 | 4096 | 2600 |
| 11 | 0xb | c0-0c0s2n3 | service | xt | (service) | 2 | 8000 | 4096 | 2600 |
| 12 | 0xc | c0-0c0s3n0 | service | xt | (service) | 2 | 8000 | 4096 | 2600 |
| 15 | 0xf | c0-0c0s3n3 | service | xt | (service) | 2 | 8000 | 4096 | 2600 |
| 16 | 0x10 | c0-0c0s4n0 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 17 | 0x11 | c0-0c0s4n1 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 18 | 0x12 | c0-0c0s4n2 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 19 | 0x13 | c0-0c0s4n3 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| ..... | | | | | | | | | |
| 2520 | 0x9d8 | c9-1c2s6n0 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 2521 | 0x9d9 | c9-1c2s6n1 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 2522 | 0x9da | c9-1c2s6n2 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 2523 | 0x9db | c9-1c2s6n3 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 2524 | 0x9dc | c9-1c2s7n0 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 2525 | 0x9dd | c9-1c2s7n1 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 2526 | 0x9de | c9-1c2s7n2 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |
| 2527 | 0x9df | c9-1c2s7n3 | compute | xt | CNL | 8 | 16000 | 4096 | 2400 |



## xtnodestat

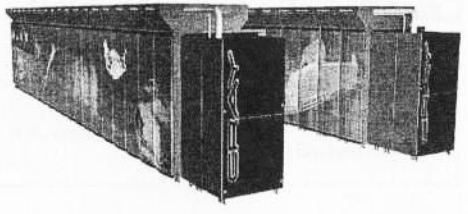


## Повестка дня

- Обзор среды программирования
  - Модули
- Компиляторы
  - PGI, Cray, GNU(, Intel, Pathscale)
- Библиотеки
- Программный анализ
- MPI взаимодействие
- Запуск приложений



# Программная среда



# Обзор



## Обзор программной среды Cray XE6

- Доступно разнообразие компиляторов
  - PGI, Cray, GNU(, Intel, Pathscale)
- Оптимизированные библиотеки
  - 64 bit AMD Core Math library (ACML)
  - Scilib: Level 1,2,3 of BLAS, LAPACK, FFT, ScaLAPACK, BLACS
  - FFTW, netCDF, HDF5, PETSc
  - MPI-2 передача сообщений библиотеки для связи между узлами
  - SHMEM односторонняя связь библиотеки
  - PGAS : CAF и UPC
- Aprun команда для запуска работы; похож на команду mpirun
- Крутящий момент пакетной системы
- Эффективность инструментов: CrayPat, Apprentice2
- Отладчик: ddt

## Программная среда Cray XE6 - SIMPLE

• Редактировать и компилировать программы (не требуется указывать включаемых файлов и библиотек)

```
$ vi mysrc.f90
$ ftn -o myexe mysrc.f90
```

• Редактирование пакетного job-файла (myjob.job) (SLURM batch)

```
#PBS -N myjob
#PBS -lmppwidth=240
aprun -n 240 ./myexe
```

• Запуск на выполнение

```
$ qsufo myjob.job
```



## Модульная панель инструментов Cray XE6

• Как мы можем получить соответствующий компилятор и библиотеки для работы?

• модульные инструменты, используемые на XE6 для обработки различных версии пакетов (компилятор, средства,...):

**например: модуль загрузки compilerl**

**например: модуль подкачки compilerl compiler2**

**например: модуль загрузки perftools**

• наблюдение за изменением PATH, MANPATH, LM_LICENSE_FILE,.... среды.

• пользователи не имеют прав устанавливать те переменные окружения в их файлач запуска оболочки, makefile'x,....

• не усложняется гибкость для поддержки других версии пакета



## Cray XE6 PE: список модулей

```
hpcander@xe601:~> module list
Currently Loaded Modulefiles:
  1) modules
  2) nodestat/2.2-1.0301.22648.3.4.gem
  3) sdb/1.0-1.0301.22744.3.46.gem
  4) MySQL/5.0.64-1.0301.2899.20.2.gem
  5) lustre-cray_gem_s/1.8.2_2.6.27.48_0.1.1_1.0301.5522.8.1-1.0301.23669.3.42
  6) udreg/1.3-1.0301.2236.3.7.gem
  7) ugni/2.0-1.0301.2365.3.7.gem
  8) gni-headers/2.0-1.0301.2497.4.1.gem
  9) dmapp/2.2-1.0301.2594.5.1.gem
 10) xpmem/0.1-2.0301.22550.3.7.gem
 11) Base-opts/1.0.2-1.0301.21771.3.3.gem
 12) xtpe-network-gemini
 13) pgi/10.9.0
 14) xt-libsci/10.4.8
 15) pmi/1.0-1.0000.7901.22.1.gem
 16) xt-mpt/5.1.1
 17) xt-asyncpe/4.4
 18) PrgEnv-pgi/3.1.37E
 19) moab/5.4.1
 20) torque/2.5.2
 21) ws/1.0
hpcander@xe601:~>
```



## Cray XE6 PE: module avail (доступные модули)

```
> module avail cce

cce/7.2.4          cce/7.2.5          cce/7.2.6          cce/7.2.7
cce/7.2.8(default) cce/7.3.0.145
```

• Какие модули доступны? ( вывод запрещен из фазы 0)

• "module avail" – выведет список всех доступных модулей

## Cray XE6 PE: отображение модулей

```
hpcander@xe601:~> module show PrgEnv-pgi
-------------------------------------------------------------------
/opt/modulefiles/PrgEnv-pgi/3.1.37E:

conflict        PrgEnv
conflict        PrgEnv-xl
conflict        PrgEnv-x2
conflict        PrgEnv-gnu
conflict        PrgEnv-pathscale
conflict        PrgEnv-intel
conflict        PrgEnv-cray
module          unload pgi
module          unload xt-totalview
module          unload xt-libsci
module          unload pmi
module          unload xt-mpt
module          unload xt-asyncpe
setenv          PE_ENV PGI
setenv          XTOS_VERSION 3.1.37E
setenv          XTPE_COMPILE_TARGET linux
prepend-path    PE_PRODUCT_LIST PGI
module          load pgi
module          load xt-libsci
module          load pmi
module          load xt-mpt
module          load xt-asyncpe
setenv          CRAY_PRGENVPGI loaded
-------------------------------------------------------------------

hpcander@xe601:~>
```



## Необходимые модульные команды

- Базовые: загрузка PGI компилятора и Magny-Cours спецификации

  **module load PrgEnv-pgi**

  **module load xtpe-mc8**
- Изменение среды (GNU)

  **module swap PrgEnv-pgi PrgEnv-gnu**
- Загрузка среды Cray и использование специальной версии компилятора

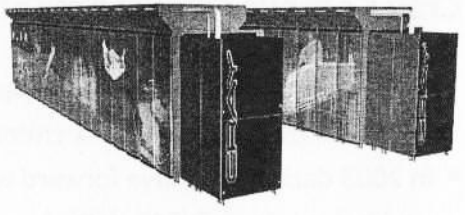
  **module swap PrgEnv-pgi PrgEnv-cray**

  **module swap cce cce/7.3.0.145**
- Только загрузка среды MPICH2

  **module unload xt-mpt**

  **module load xt-mpich2**





# Компиляторы



## Драйвера компиляторов для создания исполняемых CLE-файлов

- Когда PrgEnv загружается драйвера компилятора также загружаются
  - драйвера компилятора заботятся о загрузке соответствующих библиотек (-lmpich, -lsci, -lacml, -lpapi)
- Доступны драйверы (в том числе для связи MPI приложений):
  - Fortran 90/95 программы: ftn
  - Fortran 77 программы : f77
  - C программы: cc
  - C++ программы: CC
- Перекрестная компиляция среды
  - Компиляция на узле службы Linux
  - Создание исполняемого файла для CLE вычислительного узла
  - Не используйте стандартные имена компилятора (pgf90, GCC, mpicc,...) если Вы не хотите исполняемый файл Linux для служебного узла

## Получение хороших параметров компилятора и лучших библиотек

- Необходимо загрузить специальный модуль для создания кода, оптимизированный для AMD Magny-Cours процессорного модуля нагрузки xtpe-MC8 (или xtpe-mcl2)
- Установить важные флаги производительности компилятора (arch,...)
- Загрузить библиотеки лучшей производительности
  - Обратите внимание, что изменения в TNE среде обнаруживаются не всегда ‚module show'



## Программная среда PGI (PrgEnv-pgi)

- **Обзор опций**
  - -Mlist создает файл листинга
  - -Minfo информация о выполнении оптимизации
  - -Mneginfo почему некоторые оптимизации не выполняются
- **Опции препроцессора**
  - -Mpreprocess запускает препроцессор Fortran файлов
- **Опции оптимизации**
  - -fast выбирает оптимально флаги для целевой платформы
  - -Mipa=fast,inline межроцедурный анализ
  - -Minline=levels:n номер уровня инлайнинга



## Программная среда PGI

- **Языковые опции**
- -Mfree process Fortran source using freeform specifications
- -Mnomain useful for using the ftn driver to link programs with the main program written in C or C++ and one or more subroutines written in Fortran
- -i8, -r8 treat INTEGER and REAL variables as 64-bit
- -Mbyteswapio big-endian files in Fortran; XE6 is little endian
- **Parallelization Options**
  - -mp recognize OpenMP directives
  - -Mconcur automatic parallelization

man pages: pgf90, pgcc, pgCC
PGI User's Guide (Chapter 2) http://www.pgroup.com/doc/pgiug.pdf



## Cray Compiler Environment (CCE): (PrgEnv-cray)

- Cray has a long tradition of high performance compilers
  - Vectorization for vector architectures
- In 2008 decided to move forward with Cray X86 compiler
  - * Vector is back ? (SSE, AVX...)
  - * CCE 7.0 released in December 2008
  - CCE 7.3 about to be released
- Still young but interesting...
  - FORTRAN 2008 Coarray support
  - Strongly supported by Cray