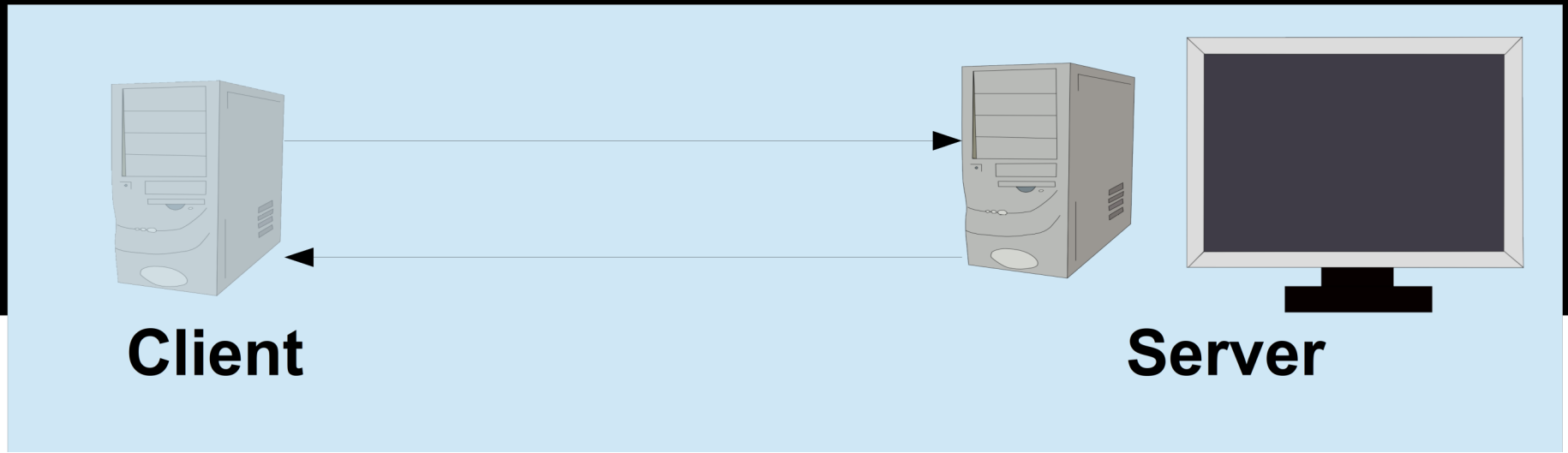
Курс "Администрирование суперкомпьютеров"

Жуматий С.А.

X-server



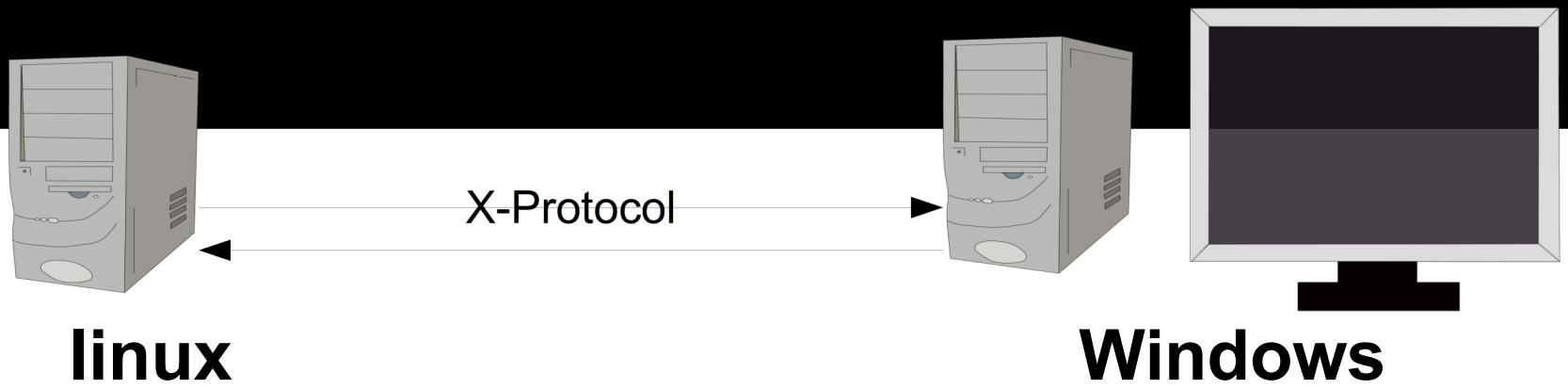
X-server

xauth — программа для управления доступом

ssh -X / -Y - «пробросить» X-соединение.

startx — запустить X-сервер

X-server



Управление задачами

Задачи

- Torque
- LSF
- LoadLeveler
- Slurm
- Cleo

Управление задачами

- Torque

- Наследник OpenPBS
- Стандарт для многих систем

- LSF

- Коммерческий продукт
- Наиболее развитый инструмент

Управление задачами

- Slurm

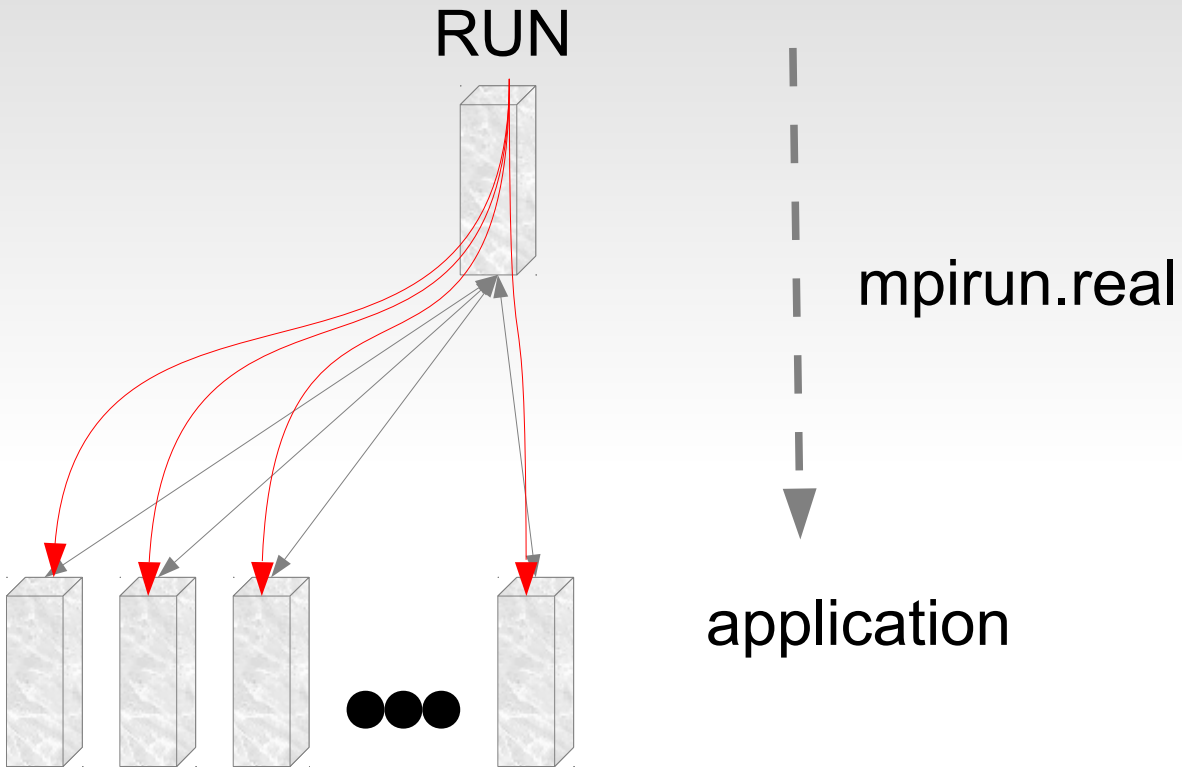
- Разработка Livermore Computing Center
- Встроенная поддержка популярных MPI
- Модульность

Управление задачами

- Cleo

- поддержка большинства вариантов MPI
- гибкая поддержка параллельных сред
- расширяемость
- контроль задач на узлах

Управление заданиями



Управление заданиями

Slurm:

sbatch/srun - поставить задачу в очередь

- -n число процессов
- -N число узлов
- -p имя раздела
- ...

Управление заданиями

squeue - просмотр задач (очереди)

-p …

-o ФОРМАТ

-u user1,user2,...

-j jobid1,jobid2,...

-n node1,node2,...

Управление заданиями

%P = partition

%i = id

%u = user

%j = jobname

%M = runtime

%e = expected end

Управление заданиями

%T = state

%D = n allocated nodes

%C = n allocated cpus

%r = reason

Пример: %10P %.7i %10u %.8j %10M %10e %10T %.5D/%5C %r

PARTITION JOBID USER TIME END_TIME STATE NODES/CPUS REASON

test2 126744 piskun mpiarray 2-21:16:42 NONE RUNNING 4/32 None

Управление заданиями

sinfo - просмотр статистики очередей

%P = partition

%a = availability (up/down/...)

%C = CPUS (alloc/idle/...)

%T = state (по строке на состояние)

%N = nodes list

Управление заданиями

sinfo

"%10P %10a %C"'

regular4 up   29712/72/2472/32256

"%10P %6D %20T"'

regular4 3714   allocated

"%10P %20C %20T %N"'

regular4 29304/0/0/29304   allocated
node1-001-[01-20,23-27,29-32], ….

Управление заданиями

scontrol - управление/просмотр BCEM

show ENTITY ID

aliases, config, daemons, frontend, job, node, partition, reservation, slurmd, step, topology, hostlist, hostnames

Управление заданиями

scontrol

scontrol update nodename=NODES State=state

Управление заданиями

Cleo

```
# tar xfvz cleo-*.tgz
# cd cleo-*
# make && make install
# ssh node-1 (cd cleo-*; make install)
```

Управление заданиями

vi /etc/cleo.conf

```
main.pe = node-1:1 node1:2 …
main.pe = node-2:1 node2:2 …
exec_line = /opt/openmpi/bin/mpirun \
  --hostfile $file \
  -n $np $task
```

Управление заданиями

vi /etc/cleo.conf

file_head = $node

coll_nodes=0

single.exec_line = ssh -t $nodes 'cd $dir; $task'

Управление заданиями

```
chgrp mpi /opt/openmpi/bin/*run*
chmod o-rwx /opt/openmpi/bin/*run*

$ mpirun -np 10 ./cpi
$ tasks
$ cleo-submit -np 10 ./cpi
```

Компиляторы

- GNU
- Intel
- Pathscale
- PGI
- cuda

Программная среда — компиляторы

- gcc/gfortran

- бесплатность
- неполная поддержка f90/f95
- не всегда хорошая производительность

Программная среда — компиляторы

- Intel

- OpenMP
- хорошая производительность (даже на AMD)
- совместимость с gcc/gfort по форматам
- относительно невысокая цена

Программная среда — компиляторы

- PathScale
- PGI

- отличная производительность
- стандарт для многих приложений
- достаточно высокая цена
- PGI — поддержка GPU (OpenACC)

Среда MPI

- mpich
- mvapich
- openmpi
- Intel
- HP/Voltair/...

Программная среда — MPI

- MPICH

- наиболее «обкатан»
- не требует установки на узлы
- для малых IP-сетей, вероятно, лучший вариант
- желательна настройка TCP-стека

- MPICH2

Программная среда — MPI

- MvaPICH/MPICH-GM

- ответвление от mpich
- официально входит в OFED(mvapich)
- оптимизация под интерконнект
- и не только (MvaPICH)
- изменённая схема запуска mvapich — mpirun не работает! Пускайте mpirun_rsh

Программная среда — MPI

- OpenMPI/LAM

- Хорошая производительность
- Официально входит в OFED
- Возможности гибкой настройки
- Почти нет документации

Программная среда — MPI

- Intel MPI

- «наследник» mpich2
- относительная независимость от среды
- поддержка отладчиков
- интегрирован в Intel Cluster Tools

OpenMPI

$wget http://www.open-

mpi.org/software/ompi/v1.6/downloads/openmpi-1.6.tar.bz2

$tar xvfj open*tbz2

$cd open*6

$./configure --prefix=/opt/openmpi

$make all install

$export PATH=$PATH:/opt/openmpi/bin

$export LD_LIBRARY_PATH=/opt/openmpi/lib64

OpenMPI

```
$ mpicc cpi.c -o cpi
```

```
$ mpirun -np 2 --host node-1,node-2 ./cpi
```

OpenMP

- GNU: -fopenmp
- Intel: -openmp
- Pathscale: -mp
- PGI: -mp

Компиляторы + MPI

Mpi-selector

--query	выбрать профиль
--list	список профилей
--set	установить профиль

- После смены профиля открыть новую сессию

Компиляторы + MPI

Вручную:

- -cc=... IMPI / OpenMPI
- MPICH_CC=... MPICH / MVAPICH / IMPI
- I_MPI_CC=... IMPI
- CC=... MPICH / MVAPICH
- OMPI_CC=... OpenMPI

Программная среда — прочее

- библиотеки
  - BLAS (MKL/APL, Atlas, GotoBLAS,...)
  - LAPACK/ScaLAPAC
  - PETSc
  - FFTW
  - прочие

Программная среда — прочее

- дополнительные средства

- Intel Threading Building Blocks
- Intel Integrated Performance Primitives
- GNU Octave (Mathlab) — с поддержкой MPI
- R, Maxima, ...
- Параллельные языки

Программная среда — прочее

- отладчики

- GNU db — только не параллельный код...
- Intel db (в комплекте с компилятором)
- Allinea DDT
- Vtune Amplifyer
- Trace Analyzer and Collector
- Dimemas, Scalasca, Extrae/Paraver
- Vampire Trace

Программная среда — прочее

- StarCD
- LS-Dyna
- Fluent
- CFX
- FlowVision
- ...

Перед покупкой поговорите с тех. специалистом компании.
«С коммерческим софтом проблем нет и поддержка там идеальна» - это миф.

Программная среда — мониторинг

-Мониторинг

- штатные SNMP-обработчики
- ganglia
- Nagios
- Zabbix
- collectd
- AntMon

Программная среда — мониторинг

- ganglia

- Очень наглядное состояние
- Нет привязки к задачам
- Ограниченное число сенсоров
- Нет реагирования

Программная среда — мониторинг

- Nagios, zabbix

- Ориентация на сервисы
- Гибкость в настройке
- Возможность реагировать на события
- Zabbix - web-интерейс

Программная среда — мониторинг

- collectd

- Гибкость в настройке
- Большое число модулей
- Базовая возможность реакции на события
- Нет управления полученными данными кроме как сохранение в rrd/log

Прочее, но не последнее

- NTP
- Backup
- UPS (NUT)
- Support
- ...

Google.com

Yandex.ru

Parallel.ru

Forum.parallel.ru

serg@parallel.ru