Московский государственный университет имени М.В.Ломоносова

Научно-исследовательский вычислительный центр

Администрирование суперкомпьютеров

С.А.Жуматий

в.н.с. НИВЦ МГУ, к.ф.-м.н.

serg@parallel.ru

Принципы построения суперкомпьютеров

А нужно ли строить?

- ***Готовые пакеты или собственные задачи?***
- ***Оценки масштабируемости?***
- ***Способ работы с системой?***
- ***Требования к оперативности обработки данных?***
- ***Бюджет?***
- ***Сроки расчётов?***

Альтернатива 1: Аренда времени

- ✔ ***Минимальные начальные затраты***
- ✔ ***Экономичность***
- ✔ ***Нулевая стоимость обслуживания***
- ✔ ***Не требуется новый персонал***
- ✘ ***Отсутствие гарантии сроков***
- ✘ ***Потенциальные проблемы с ПО***
- ✘ ***Передача данных***
- ✘ ***Безопасность***

Альтернатива 2: Аренда площадки

- ✔ *Полный контроль оборудования*
- ✔ *Гарантии инфраструктуры*
- ✘ *Стоимость аренды*
- ✘ *Требования к специалистам*
- ✘ *Необходимость поездок*
- ✘ *Передача данных*
- ✘ *Безопасность*

Альтернатива 3: Покупка готового решения

- ✔ ***Полный контроль оборудования***
- ✔ ***Краткие сроки реализации***
- ✘ ***Стоимость покупки***
- ✘ ***Требования к специалистам***

Альтернатива 4: Смена платформы (GPGPU, FPGA, персональные суперкомпьютеры, ...)

- ✔ ***Экономия средств***
- ✔ ***Краткие сроки реализации***
- ✔ ***Низкие требования к инфраструктуре***
- ✘ ***Необходимость переписывания кода***
- ✘ ***Не для всех задач подходит***

Альтернатива 5: Собственный суперкомпьютер

- ✔ ***Полный контроль***
- ✔ ***Возможность апгрейда, модификации***
- ✘ ***Требования к инфраструктуре***
- ✘ ***Требования к специалистам***
- ✘ ***Стоимость***

О чём подумать:

- Стоимость постройки/аренды
- Стоимость владения
- Требования по специалистам
- Возможность использования ПО
- Условия удалённого доступа и передачи данных
- Условия обслуживания

Структура кластера

- Сети: управляющая, вычислительная, сервисная
- Управляющий узел
- Служебные узлы
- Вычислительные узлы
- Сетевая файловая система
- Система архивирования

Служебные узлы

- DHCP — управление адресами
- TFTP — загрузка, установка
- NFS — загрузка, root-fs
- DNS — управление сетевыми именами
- NTP — синхронизация времени
- LDAP — управление учётными записями

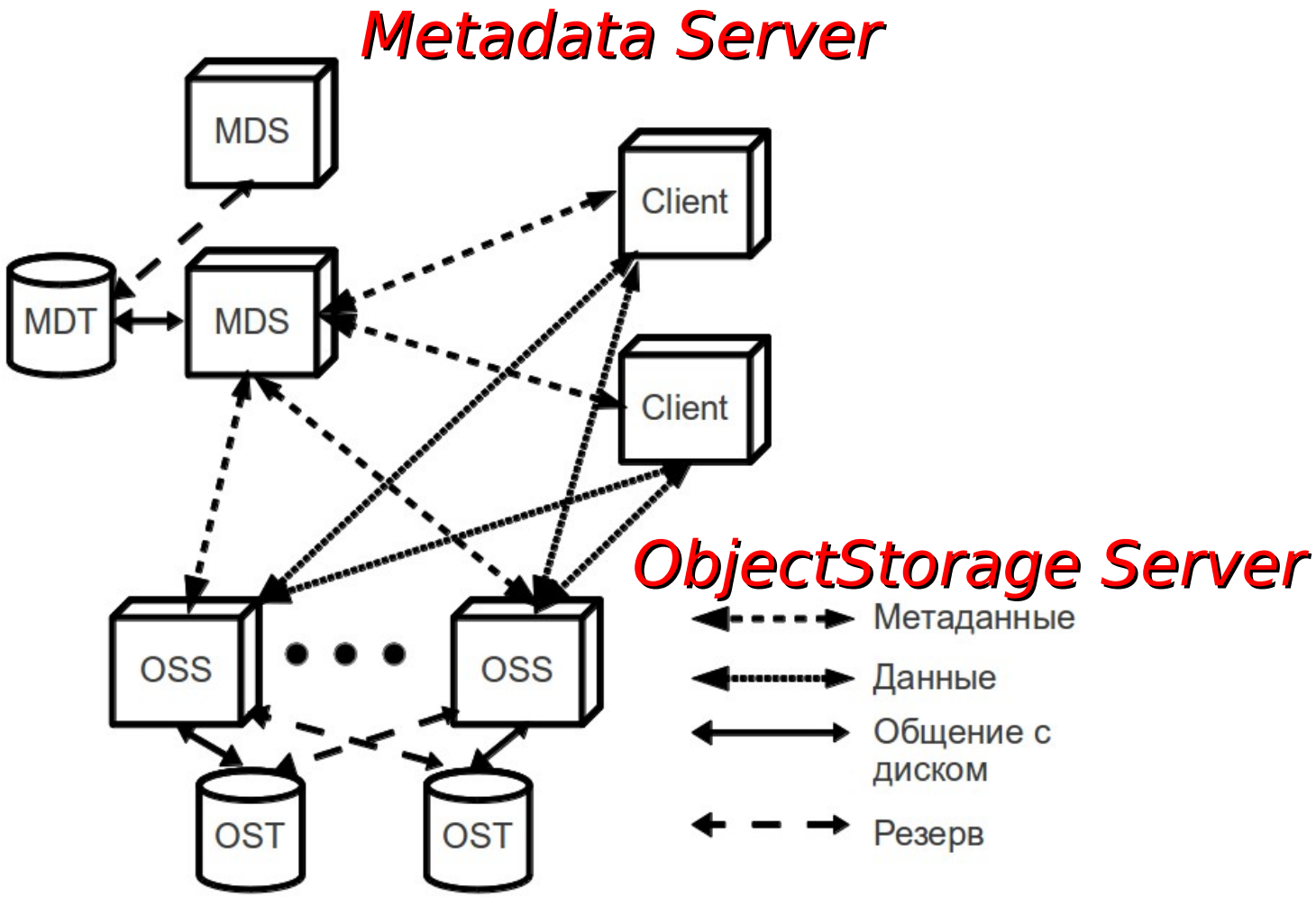
Служебные узлы

- Доступ (управляющий)
- Управление задачами
- Лицензии (flexlm)
- Мониторинг
- Статистика
- Визуализация
- Копирование данных

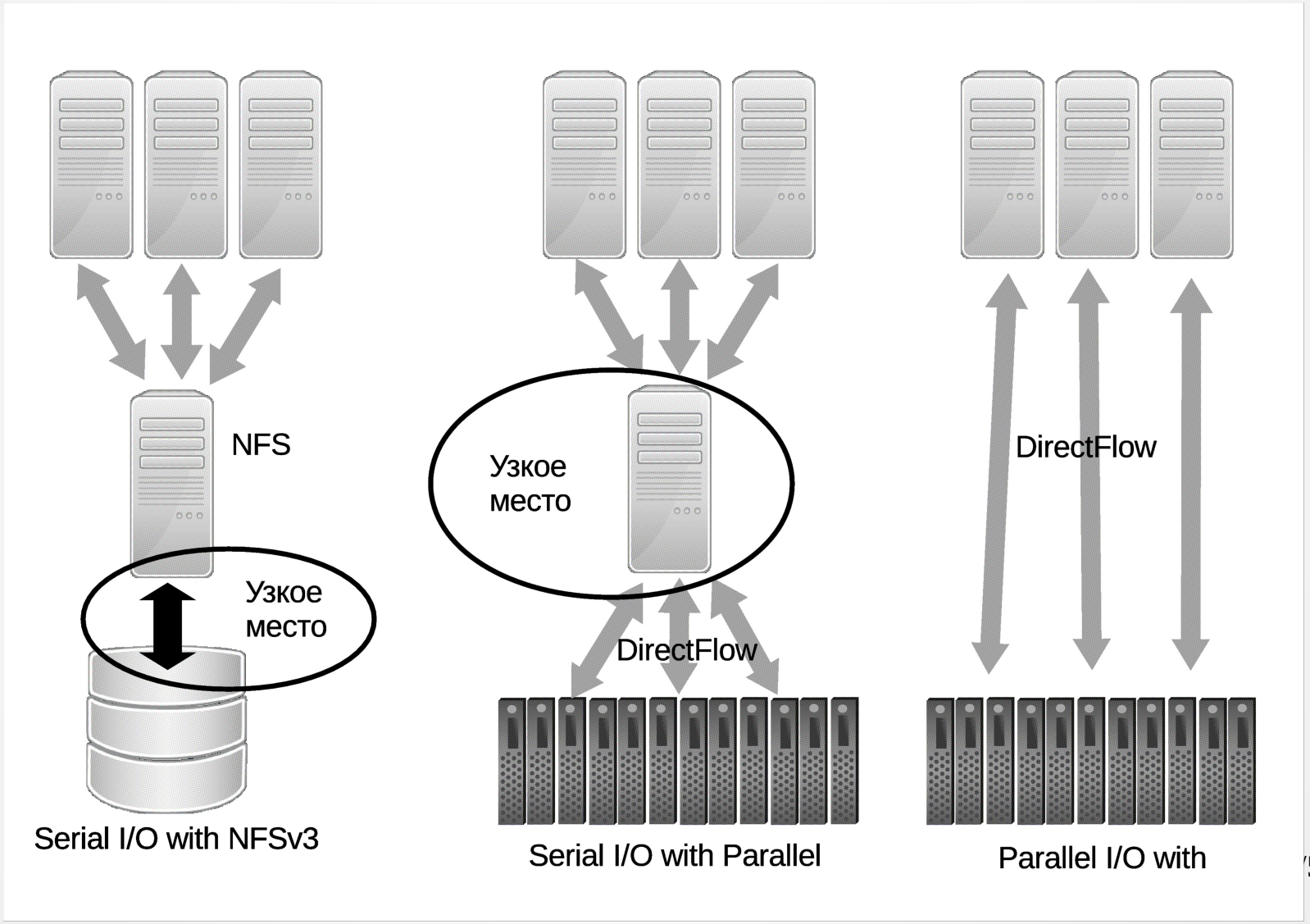
Сетевая файловая система

- NFS
- Lustre
- GPFS
- PanFS
- ...

Lustre


Metadata Server
MDS
MDT
MDS
Client
Client
ObjectStorage Server
OSS
OSS
OST
OST
Метаданные
Данные
Общение с диском
Резерв

PanFS


NFS
Узкое место
Serial I/O with NFSv3
Узкое место
DirectFlow
Serial I/O with Parallel
DirectFlow
Parallel I/O with

Установка системного ПО

- «по одному», клонирование
- Автоматизированная установка
- Готовые стеки
- XCAT

Установка системного ПО: Автоматизированная установка

RH: anaconda-ks.cfg, SuSE: autoyast.xml, Debian: preseed.file

linux **ks=http://путь-к-файлу.cfg**

linux **autoyast=http://путь-к-файлу.xml**

linux **auto preseed=http://путь-к-файлу**

Установка системного ПО: готовые стеки

- ROCKS
- PelicanHPC / Parallel Knoppics

Установка системного ПО: xCAT

- Использует anaconda/autoyast + генерация образов для загрузки по сети
- Управляет DNS, DHCP, TFTP
- Поддерживает удалённое управление питанием (через IPMI)
- Имеет встроенные «массовые» команды

Управление пользователями

- Организация доступа
- Хранение учётных записей
- Управление доступом
- Квоты

Организация доступа

- SSH + X-Forward
- VNC, RDP
- X2go, SPICE, ...

Хранение учётных записей

- passwd
- nis+
- LDAP
- Гибридный метод

Квоты

- Дисковые
- Процессорные
- Приоритеты в очереди

Управление задачами

- Torque
- SLURM
- Cleo
- LSF, LoadLeveler, ...

Управление задачами

- Torque

- Наследник OpenPBS
- Стандарт для многих систем

- LSF

- Коммерческий продукт
- Наиболее развитый инструмент

Управление задачами

- Slurm

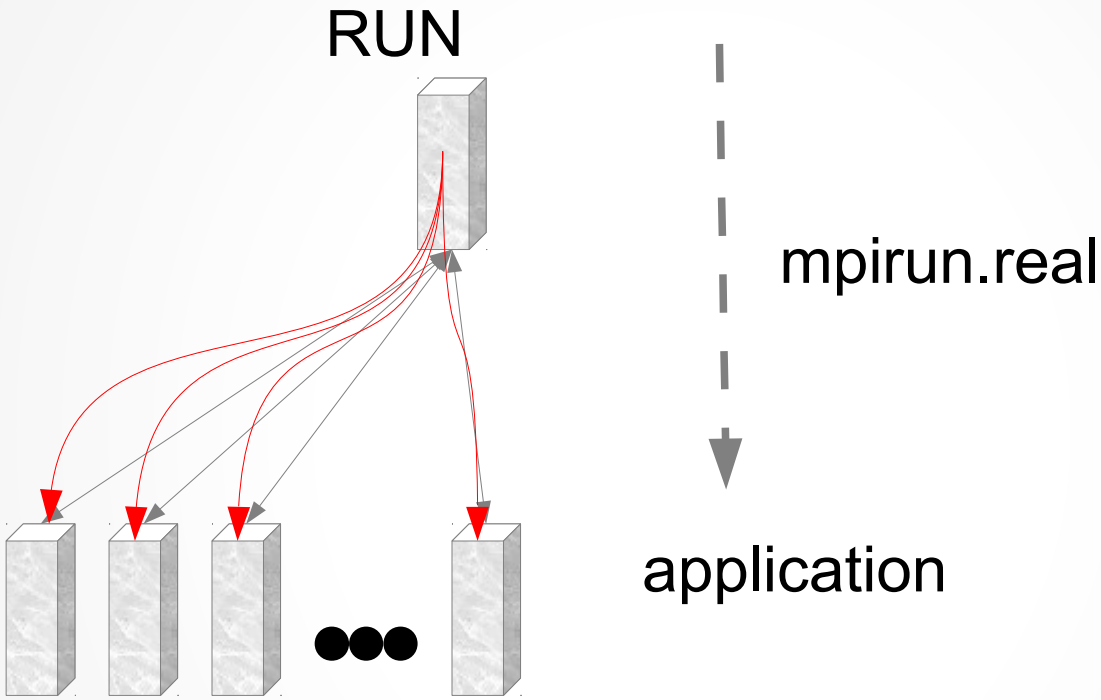
- Разработка Livermore Computing Center
- Встроенная поддержка популярных MPI
- Модульность

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 %10M %10e %10T %.5D/%5C %r

```
PARTITION    JOBID USER   TIME   END_TIME  STATE  NODES/CPUS
REASON
test2       126744 piskun 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 всем

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

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

```
$ 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

Вручную:

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

Программная среда — мониторинг

Мониторинг

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

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