Направи си сам СУперкомпютър

Д-р Христо Илиев, НИС при СУ "Св. Климент Охридски"

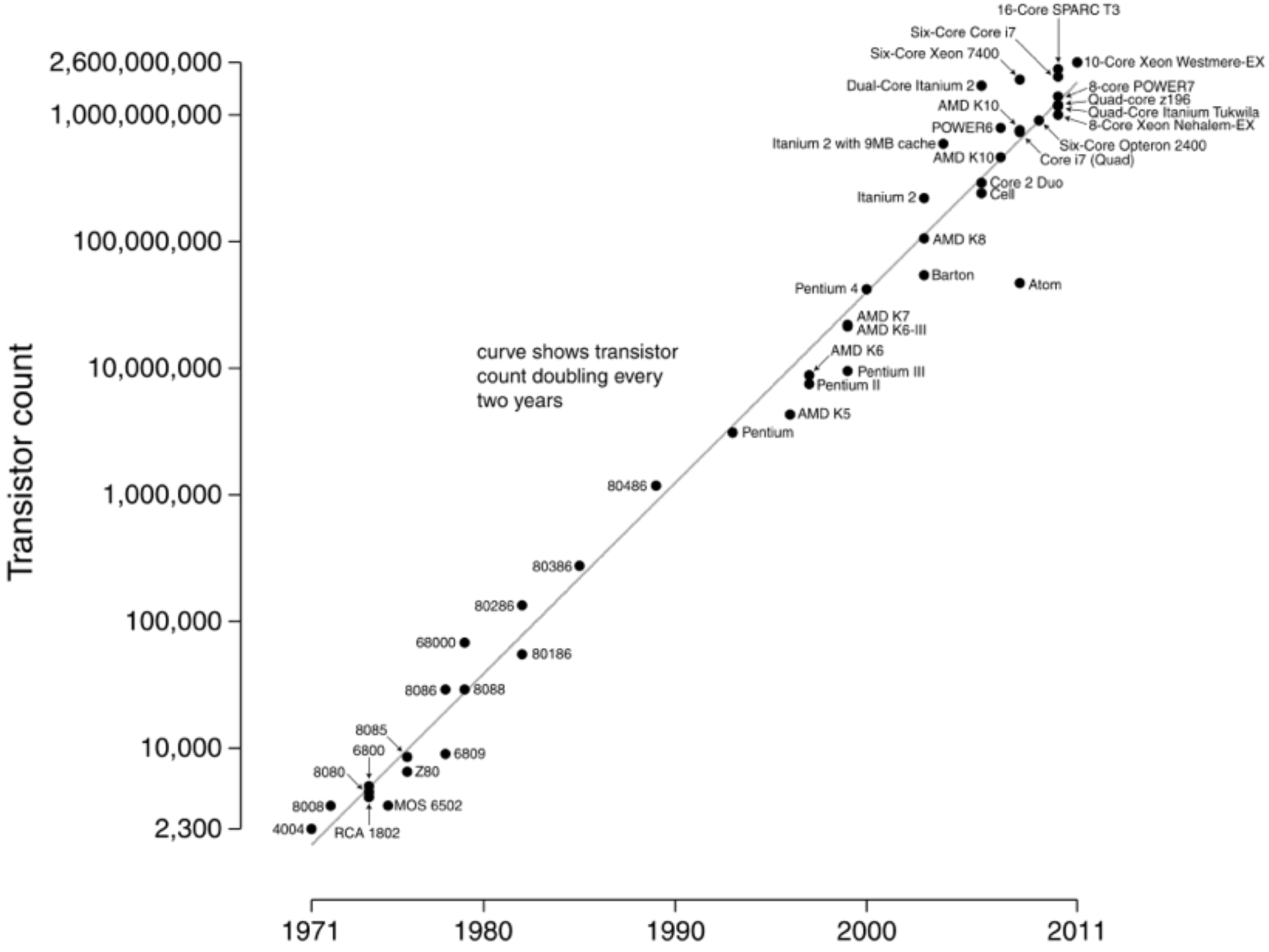
По-известни физици

- John Atanasoff – доктор по теоретична физика
- John von Neumann – доктор по математика и физика
- Edsger Dijkstra – магистър по физика
- Donald Knuth – бакалавър по физика
- Dennis Ritchie – бакалвър по физика
- Brian Kernighan – бакалавър по инж. физика
- Richard Stallman – бакалвър по физика

Суперкомпютър

- Голям
- Бърз
- Енергоемък
- Скъп
- Огромен дисков капацитет
- Паралелен

Microprocessor Transistor Counts 1971-2011 & Moore's Law



Закон на Мур

- Напасване на експерименталните наблюдения с показателна функция
- $N_{\mathrm{tran}}(t) = N_{\mathrm{tran}}(t_0) \times 2^{(t-t_0)/2}$
  (Gordon Moore, CEO Intel, 1975)
- $\mathrm{OpW}(t) = \mathrm{OpW}(t_0) \times 2^{(t-t_0)/1{,}5}$
  (David House, Intel)
- Intel упорито се опитват да поддържат законите в сила!

Модел на времето

- Динамика на флуидите и термодинамика
- Диференциални уравнения ⇒ диференчни уравнения

- Симулиран обем = площ S × височина H
- Пространствена разделителна способност ΔL
- Времева разделителна способност Δt
- операции ~ $(S \times H)/(\Delta L)^3\, T/\Delta t$

flops

- floating-point operations per second
  floating-point operations → flops/s

- IEEE 754-2008

  - единична точност (single, binary32)
    32 бита; 7 десетични знака след запетаята

  - двойна точност (double, binary64)
    64 бита; 15 десетични знака след запетаята

HPL

- Стандартен начин да си ги мерим [суперкомпютрите]
  - LINPACK тест на J. Dongara – $\mathbf{A} \cdot \mathbf{x} = \mathbf{y}$ посредством гаусова елиминация с частичен избор на водещ елемент
- HPL – паралелна MPI версия
- BLAS
- $2/3 \times N^3 + 2 \times N^2$ DP ops
- N := dim(A) = колкото позволява паметта
- Резултат в DP Gflops

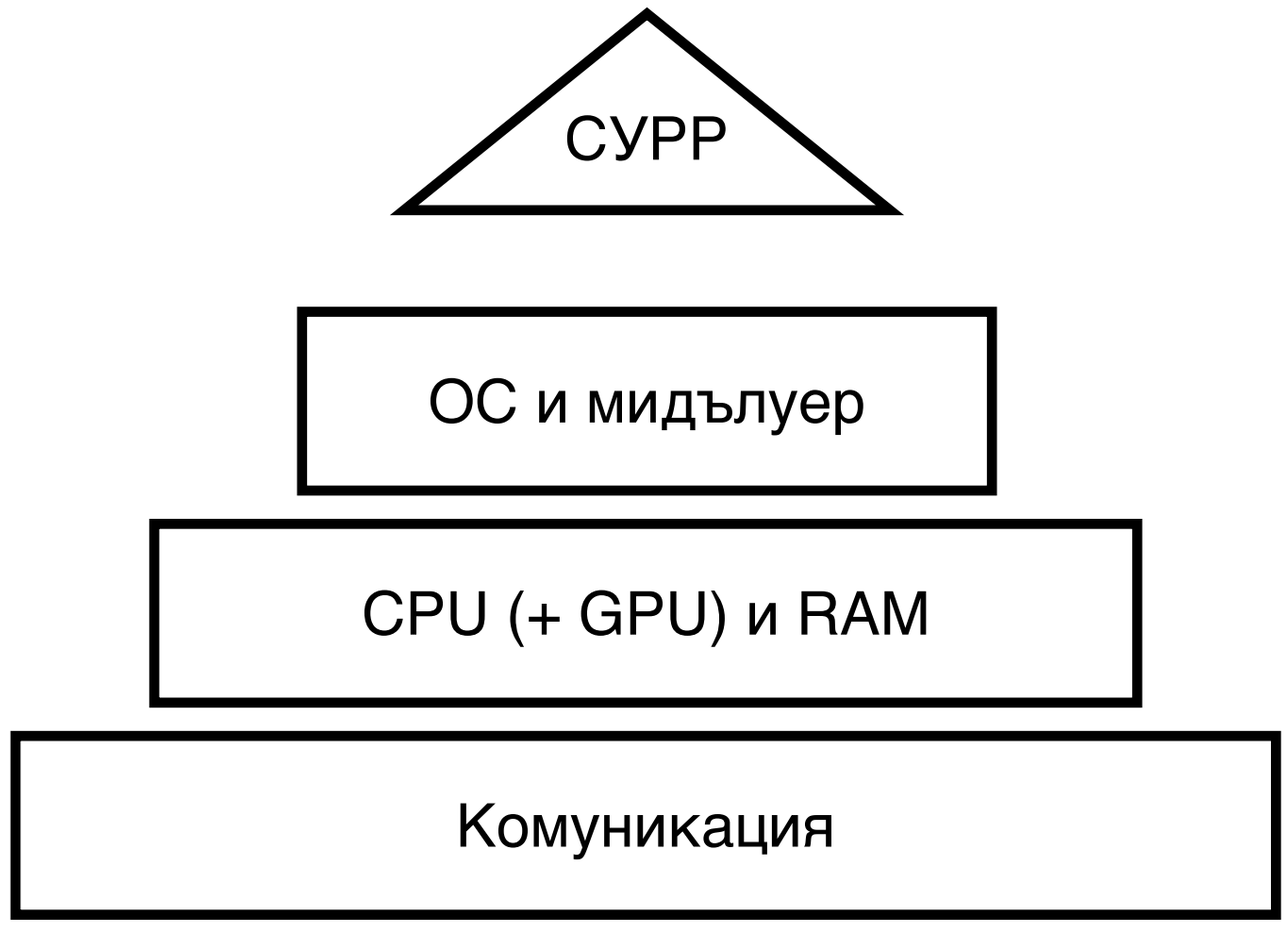
Top500

- Списък на 500-те най-бързи суперкомпютри в света
- 2 пъти годишно – юни и ноември
- $R_{peak} = N_{ядра} \times f_{такт} \times R$
  (R = 4 flops/такт на повечето съвременни процесори)
- R_{max} = максимална производителност според HPL
- N_{max} = dim(A) за постигане на Rmax
- $R_{max}/R_{peak} < 1$ – паралелна ефективност

България в Top500

- Blue Gene/P на ДАИТС (сега МТИТС)
  - 8192 PowerPC 450 ядра @ 850 MHz
  - R_{max} = 23,42 Tflops; R_{peak} = 27,85 Tflops
  - 11.2008 г. – 126 място
  - 06.2009 г. – 245 място
  - 11.2009 г. – 377 място
  - 06.2010 г. – изпада (R_{max}[#500] = 24,67 Tflops)

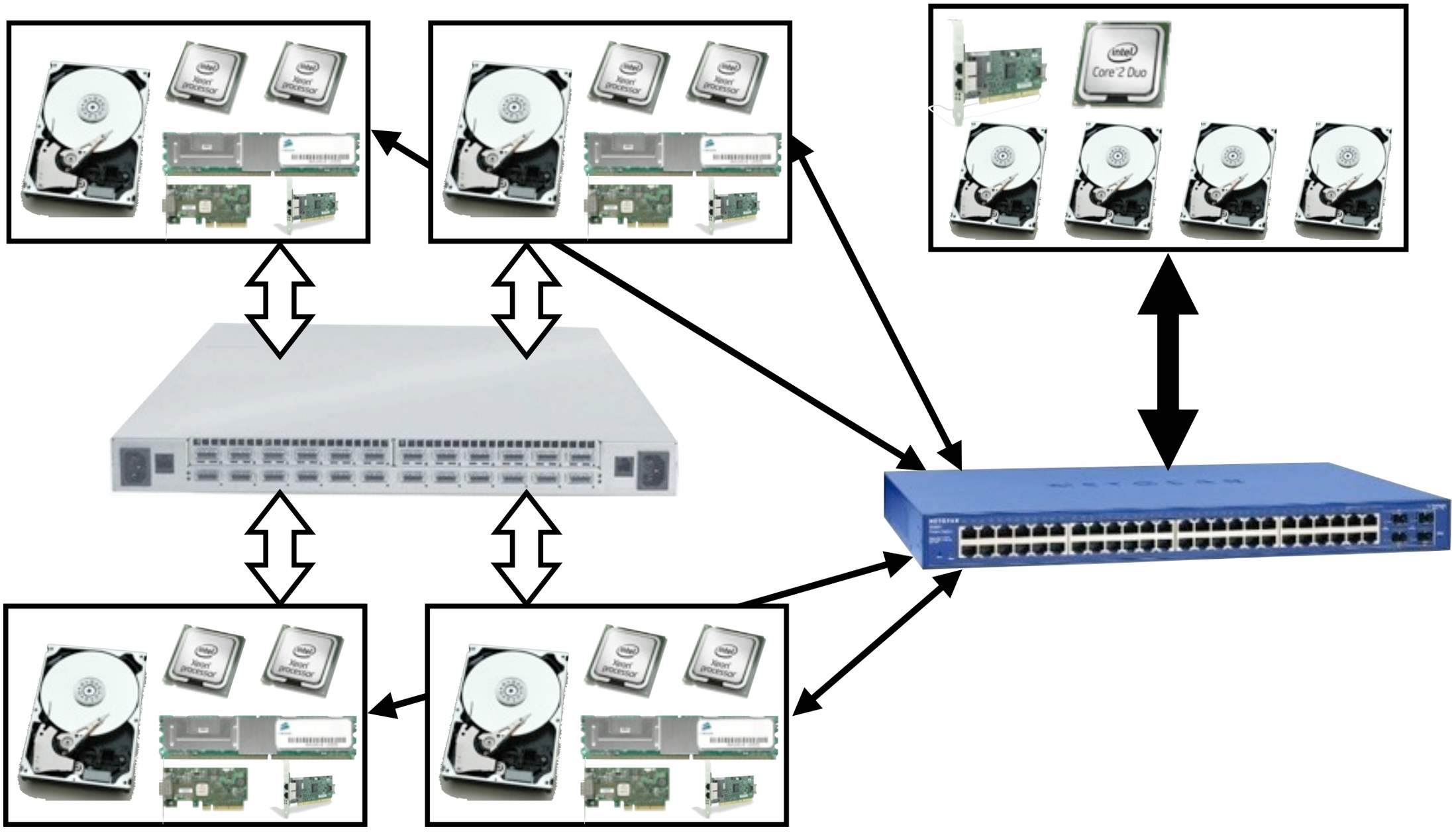
Системна йерархия



Beowulf

- Thomas Sterling и Donald Becker (NASA, 1994 г.)
- Стандартни компоненти от магазина
- Стандартна мрежова среда
- Unix-подобна ОС
- Мидълуер за паралелна обработка: MPI, PVM и др.
- Клъстери и мрежи от работни станции

Типичен клъстер



Нива на паралелизъм

- Много груб – SETI@Home, BOINC
- Груб – тривиално паралелни алгоритми
- Среден – MPI, PVM, DSM
- Фин – нишки, OpenMP, GPU
- Много фин – ILP, SIMD, GPU

Комуникация

- Обмяна на данни между процесорите
  - InfiniBand / 10 GbE
  - Myrinet
  - GigE (само за груб паралелизъм)
- Файлова система и сервизна мрежа
  - InfiniBand / GigE

InfiniBand

- Комутация на пакети (switched fabric)
- Ниска латентност за MPI ~μs
- Висока пропускателна способност
  - DDR – 20 Gbps; QDR – 40 Gbps
- Скъпи кабели :)
- OpenFabrics Enterprise Distribution (OFED)

CPU или GPU?

- Цена и енергоемкост с/у универсалност
- CPU – универсални и енергоемки
  - Intel E7-4870: <1 Gflops/W; >20 €/Gflops
- GPU – масивно паралелни и енергоефективни
  - AMD/ATI HD6950: >3 Gflops/W; <0,5 €/Gflops

CPU

- Универсални
- Паралелни и последователни алгоритми
- Голям обем памет на ядро
- Многозадачен режим
- Бавен достъп до RAM

GPU

- Масивно паралелни SIMD
- Тесен клас силно паралелни по данни алгоритми
- Приставки (ускорители) към CPU
- Малък обем собствена RAM на ядро
- Тясно място – прехвърляне на данни от/до RAM на CPU

Програмни модели

- SIMD – неявна поддръжка чрез кодовия генератор на компилатора
- Нишки – OpenMP, POSIX/Win32 нишки
- DSM – CIOMP, vSMP, Unified Parallel C
- GPU – CUDA, CAL, OpenCL, OpenMP подобия
- Предаване на съобщения – MPI, PVM

ОС

- Практически всяка ОС с мрежов стек
- Unix
  - UNICOS (Cray)
  - IRIX (SGI)
  - Solaris (Sun Microsystems)
  - Linux (IBM, SGI, Cray, Beowulf)
  - Mac OS X
    (System X на Virginia Tech, #3 в Top500 от 11.2003 г.)
- Windows 2008 HPC Server (Cray, Bull)

Мидълуер

- Предаване на съобщения
- Достъп до отдалечена памет
- Глобални операции
- Синхронизация
- Паралелен В/И
- Настройка и профилиране на паралелни приложения

Open MPI

- Реализация на MPI-1 и MPI-2
- Нов BSD лиценз
- Множество платформи и преносни среди
- Активна разработка
- http://www.open-mpi.org

Файлова система

- Общ изглед на файловата система посредством мрежово споделяне

- NFS – лесна за разгръщане, но с лоша мащабируемост

- Lustre – трудна за разгръщане, но с висока производителност и мащабируемост

Lustre

- Високомащабируема паралелна файлова система за големи клъстерни инсталации
- GPL
- CMU → CFS → Sun → Oracle → Whamcloud
- Сървър за метаданни и множество блокови хранилища
- Поддръжка на InfiniBand свързаност

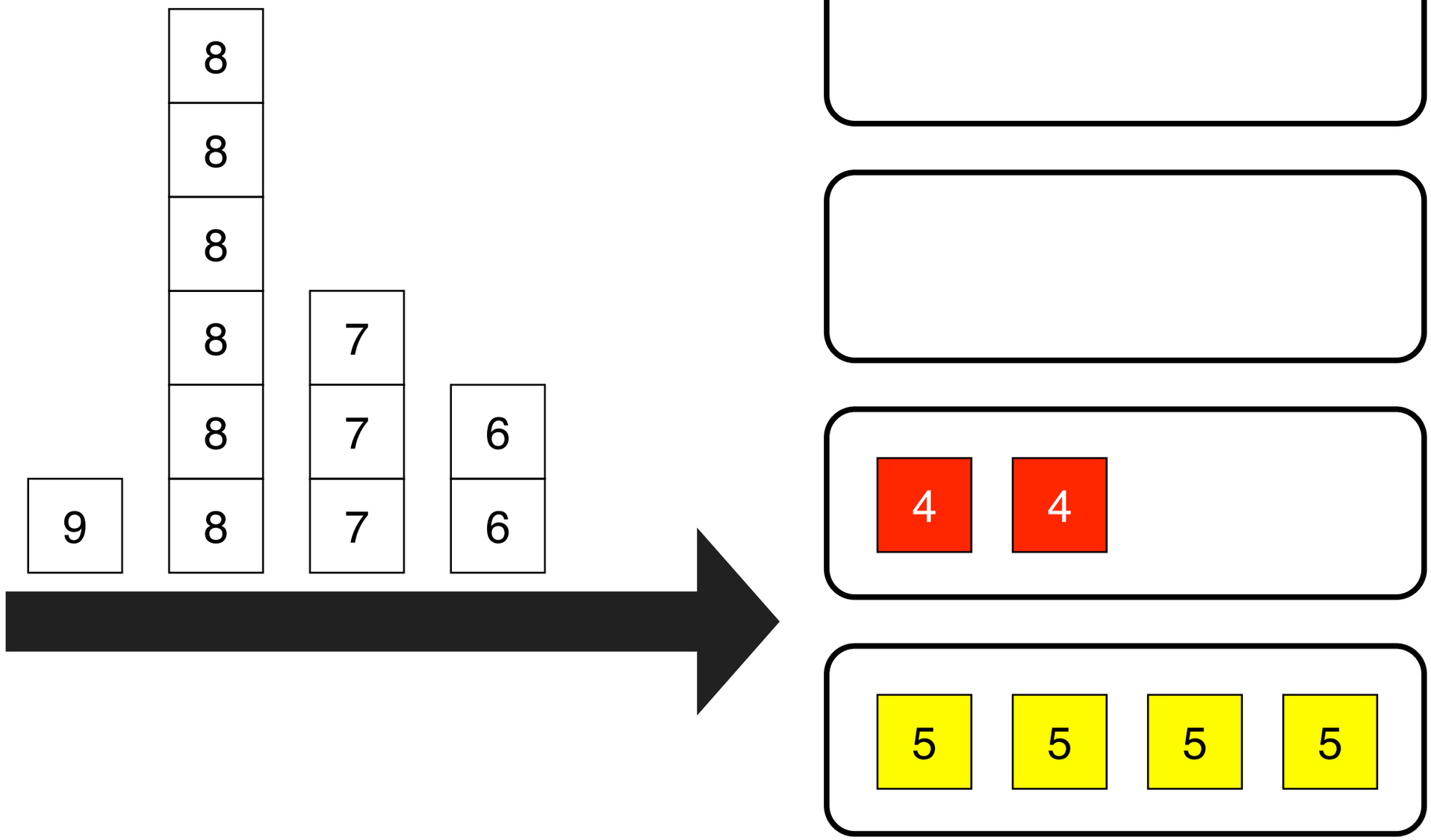
Управление на ресурсите

- Голям брой отделни възли под управление на собствена ОС (може и на различни ОС)
- Отдалечено стартиране на процеси
- Пренасочване на В/И
- Счетоводство на използваните ресурси (все някой /трябва да/ плаща)

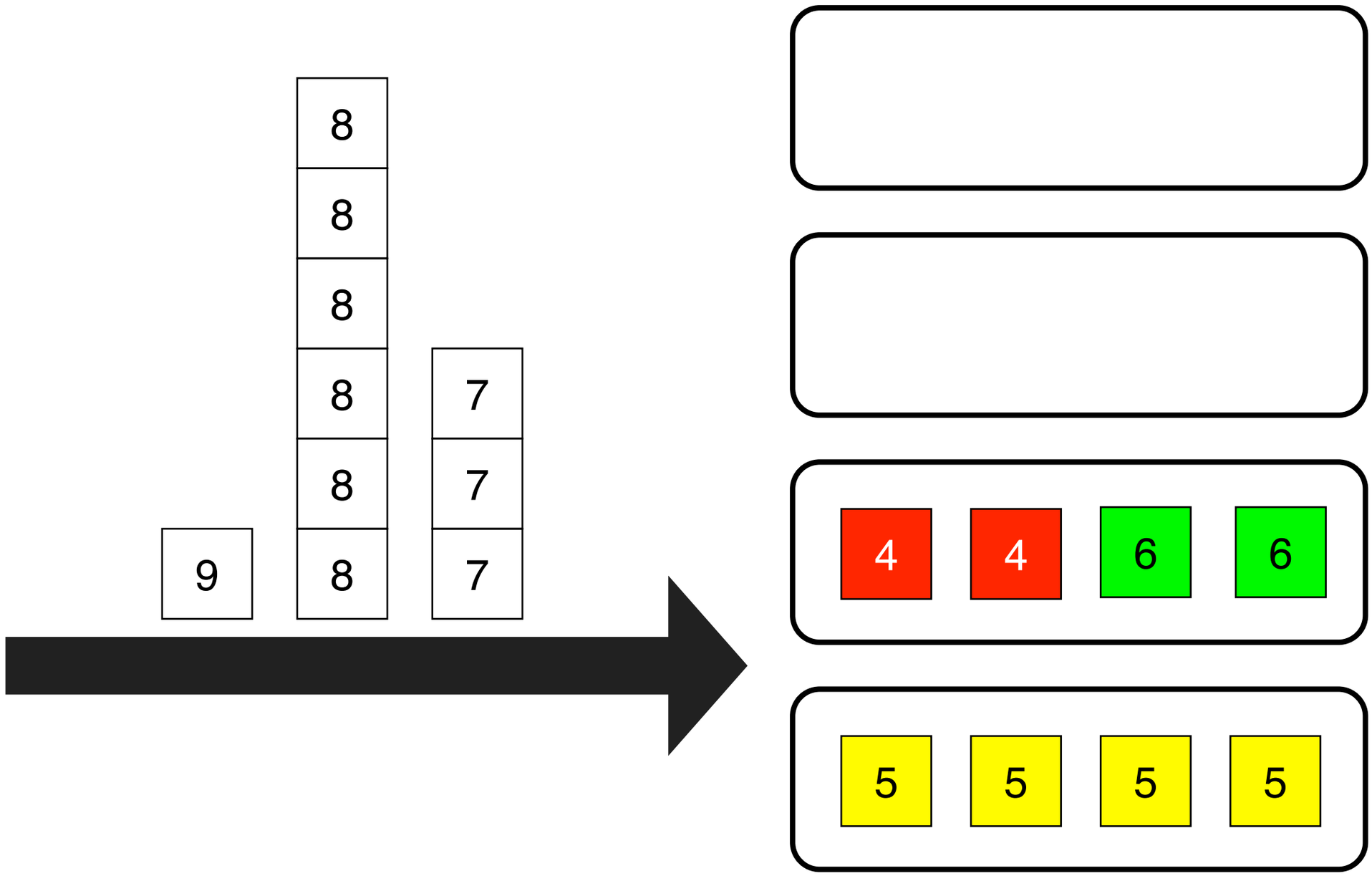
СУРР

- TORQUE + Maui Cluster Scheduler
- Open Grid Scheduler
- Simple Linux Utility for Resource Management (SLURM)
- Condor

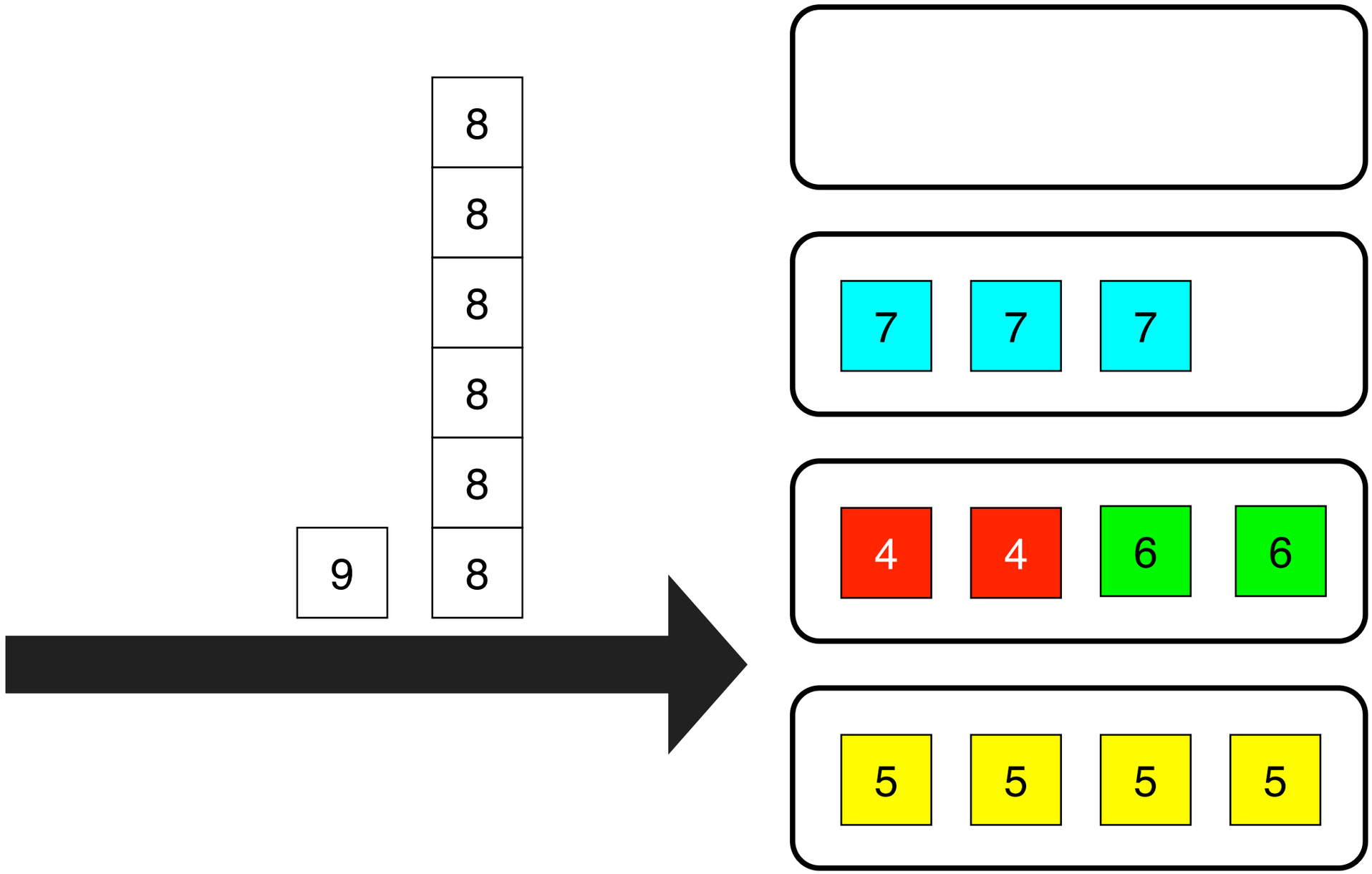
Опашки



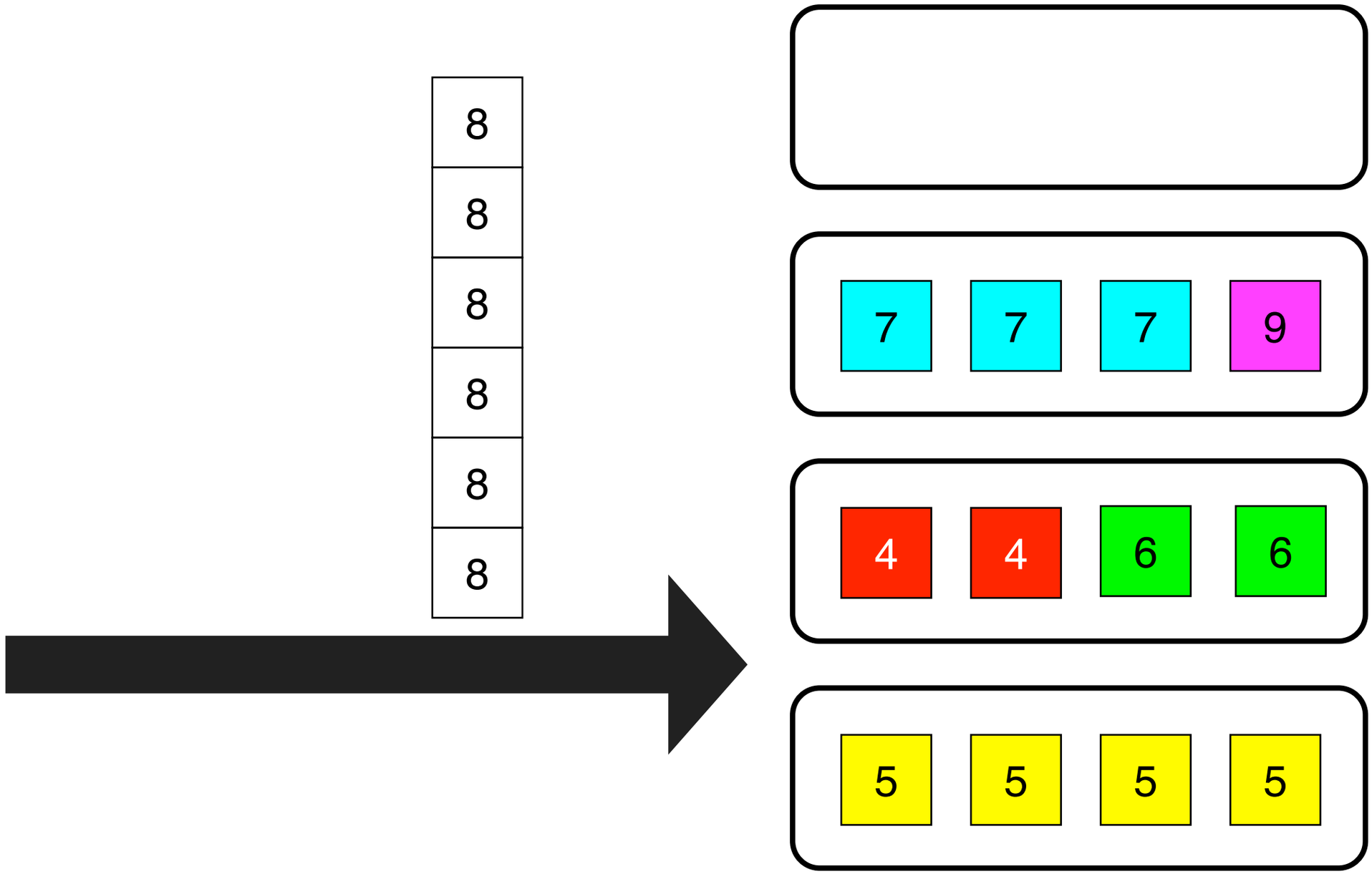
Опашки



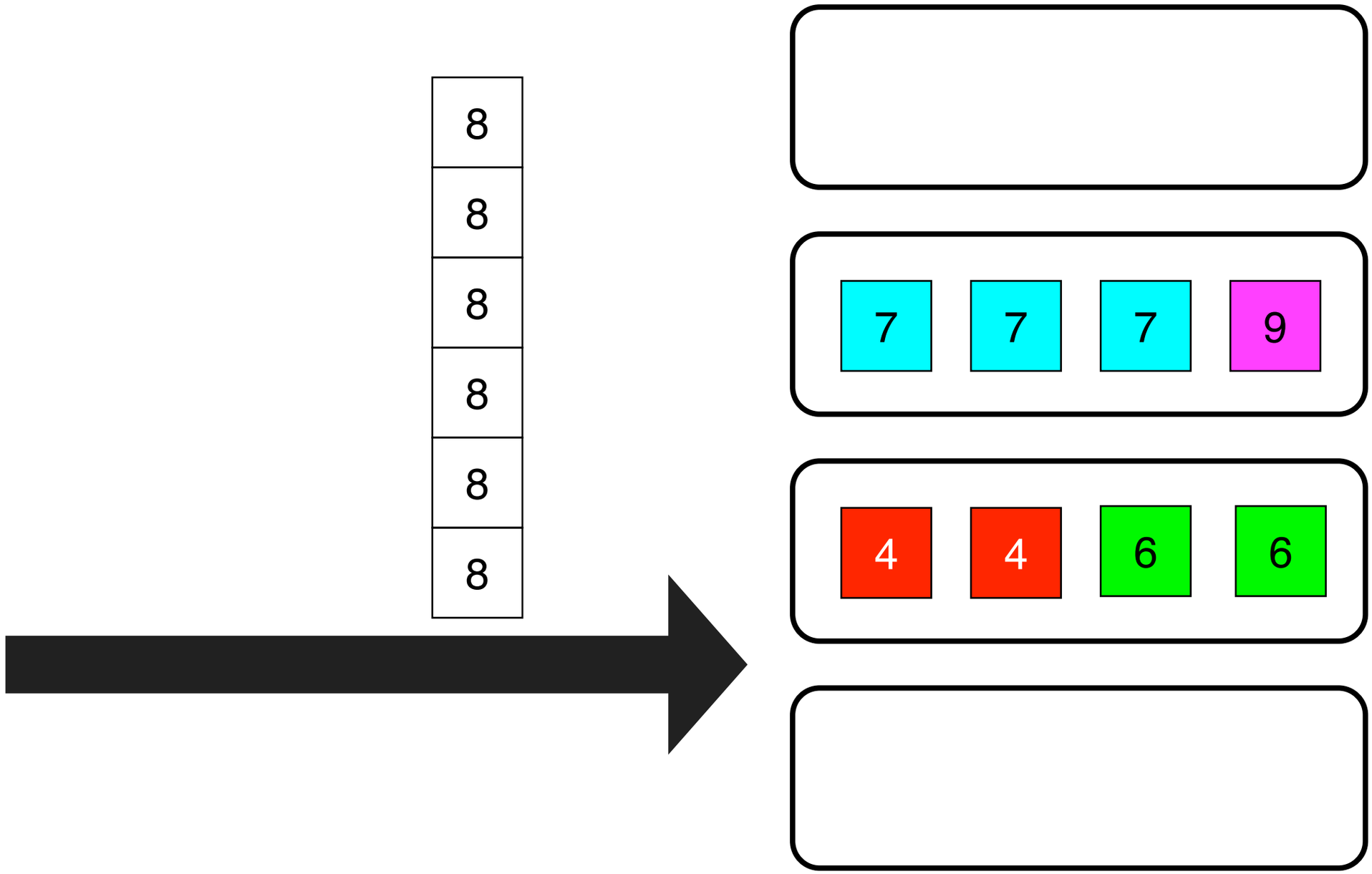
Опашки



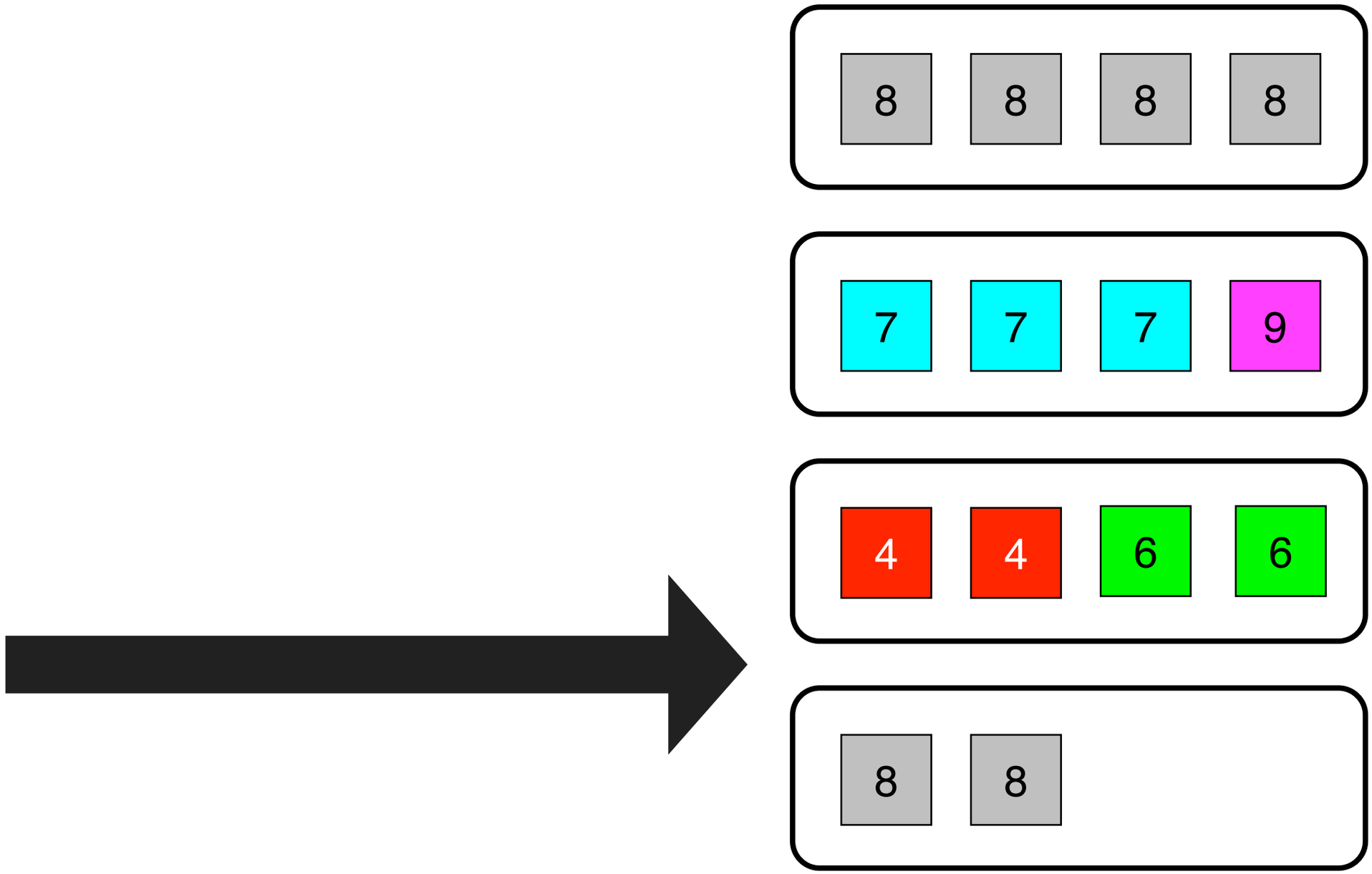
Опашки



Опашки



Опашки



Open Grid Scheduler

- Версия с отворен код на Oracle Grid Engine
- Началник-планировчик – qmaster/scheduler
- Изпълнители и пастири – execd, shepherd
- Набор от програми за мрежово взаимодействие с главния процес – qsub, qstat, qdel, qconf, qhost
- Политики за честно споделяне на ресурсите

Поддръжка

- Повече възли – по-сложна поддръжка
- Блейд модули
- Мрежово зареждане на възлите от обща инсталация
- IPMI модули за отдалечена администрация
- Наблюдаване на възлите
- xCAT за особено големи инсталации

Ganglia

- Софтуер с отворен код за наблюдение на мрежа от компютри
- Агрегация на историческа информация (rrd)
- Интеграция с gexec за отдалечено изпълнение
- http://ganglia.info/

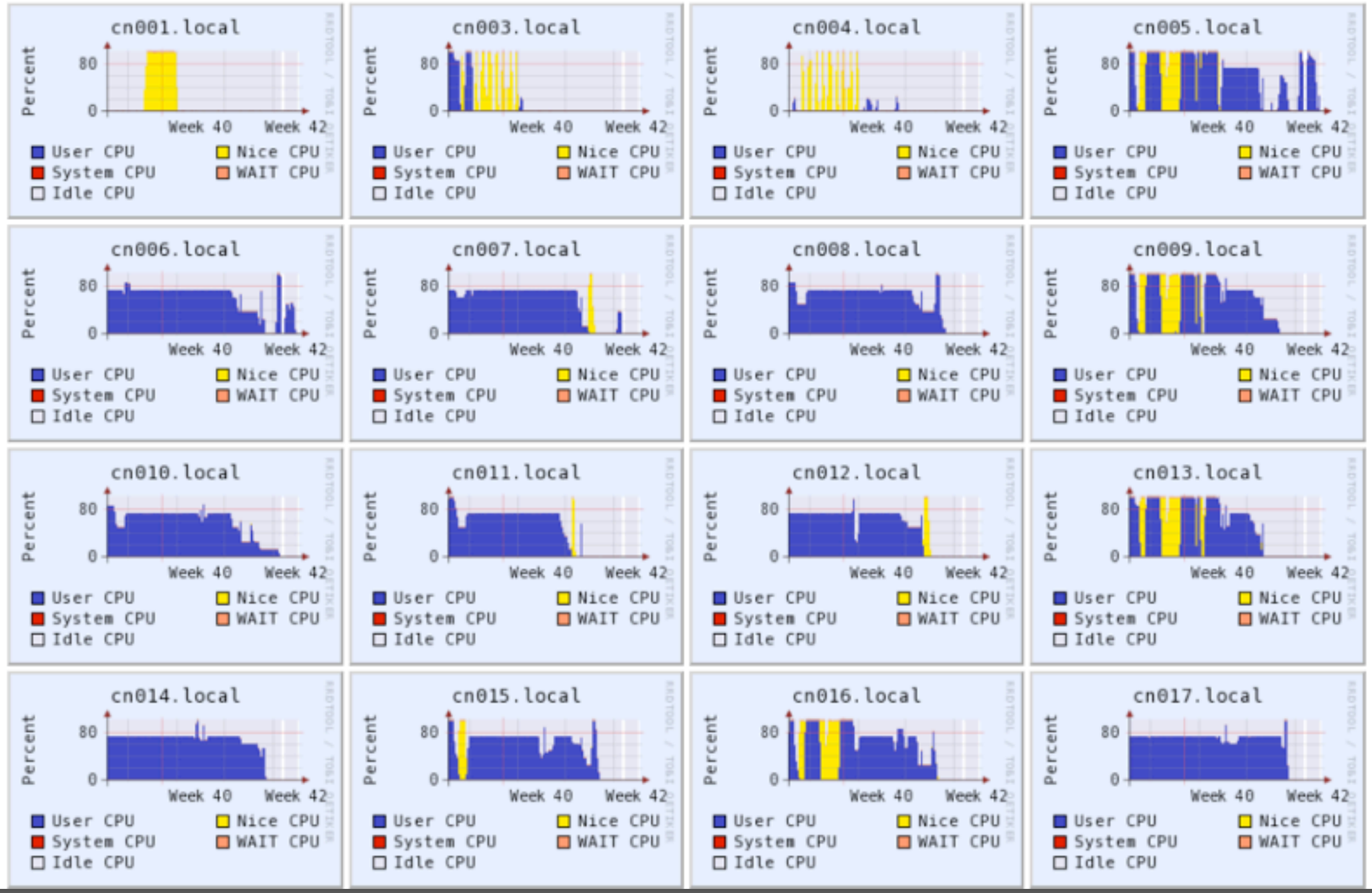
iPad 20:50 72%

physon.phys.uni-sofia.bg:8082/ganglia/?m=cpu_report&r=month&s=by%2

Google

Ganglia:: Cluster Report

Show Hosts: yes ◉ no ○ | Compute **cpu_report** last **month** sorted **by name** | Columns 4



PHYSON

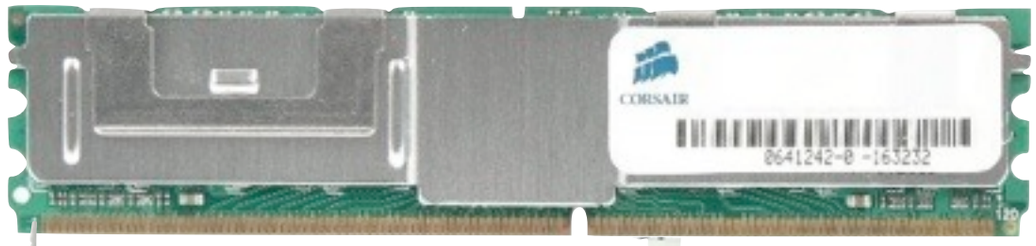
- "Розовият" клъстер на СУ
- Роден и отгледан по дог. ВУ-Ф 205/2006, поддържан по ДО 02-136/2008, ДО 02-167/2008 и ДДВУ 02-42 с НФНИ
- ~ 170 хил. лв. компютри + инфраструктура
- R_{peak} = 3,245 Tflops
- Научни пресмятания и обучение
- 13 проекта, 49 потребителя

PHYSON – възел

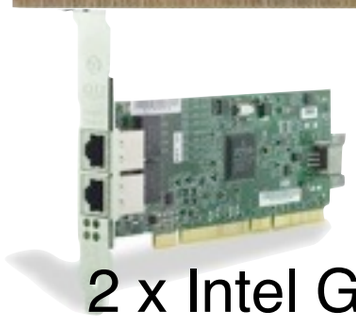
250 GB SATA2

2 бр. Intel Xeon E5420

16 GiB ECC FB-DIMM

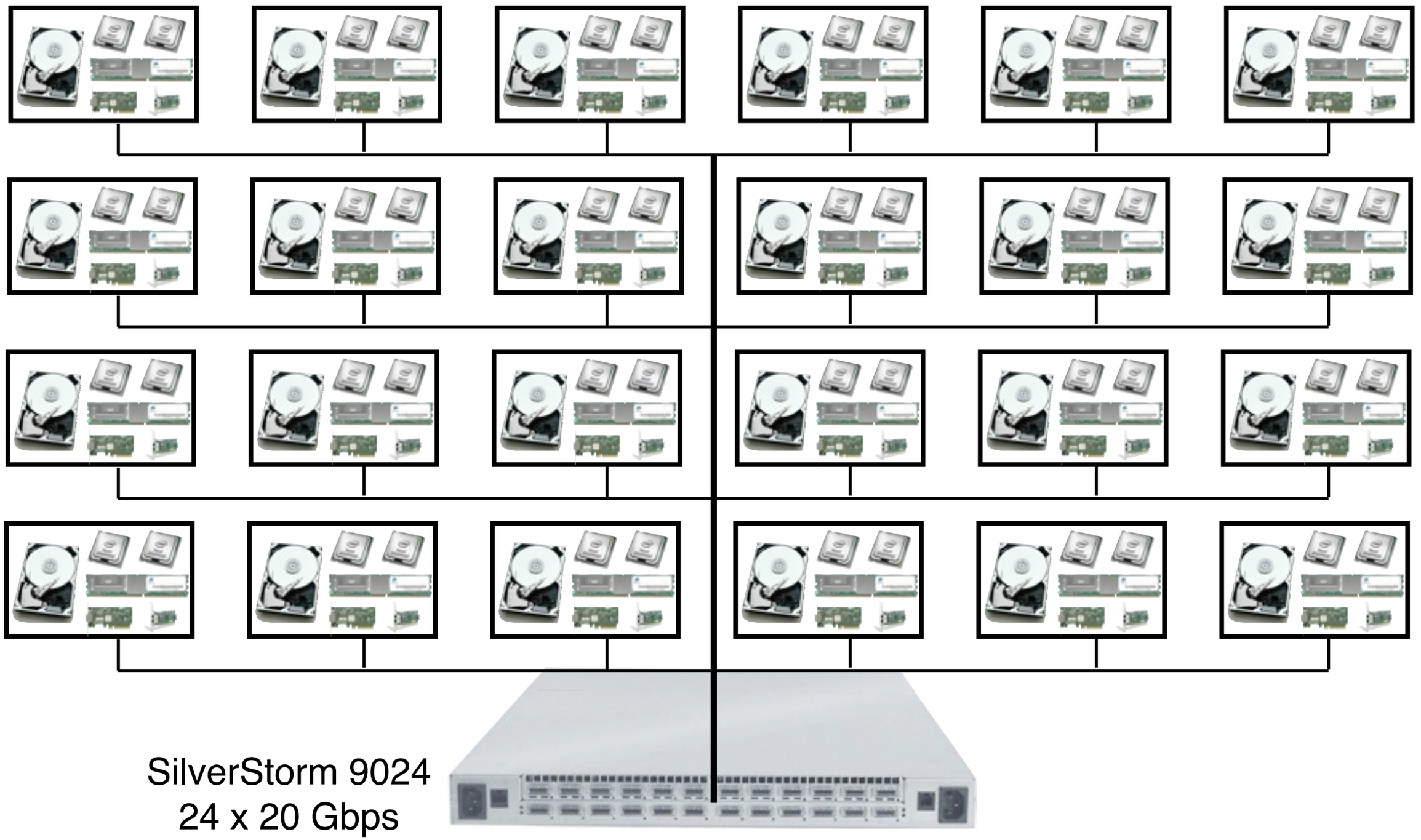


Mellanox MT25204
20 Gbps (DDR) HCA



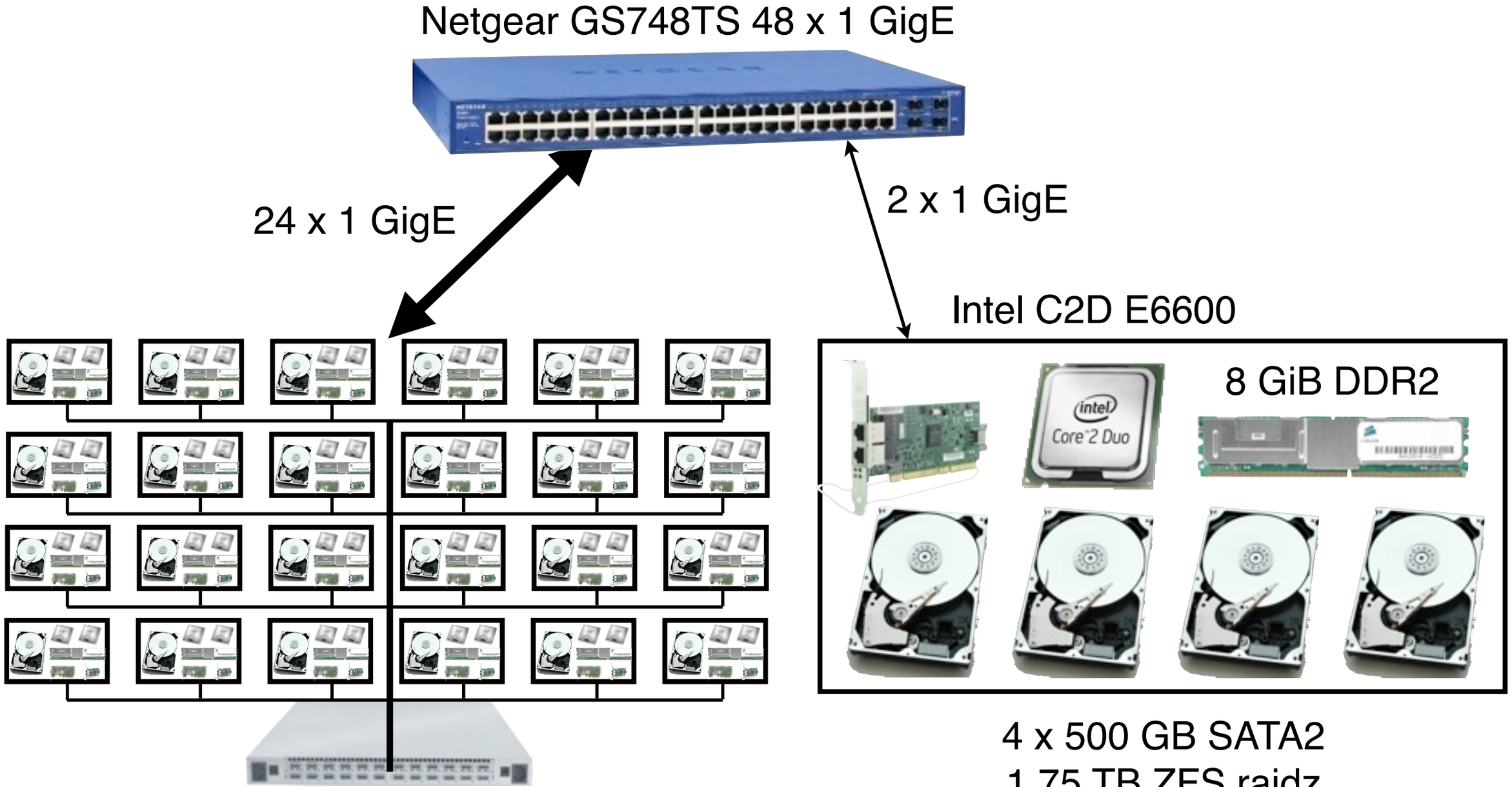
2 x Intel GigE

PHYSON – IB/MPI



SilverStorm 9024
24 x 20 Gbps

PHYSON – NFS


Netgear GS748TS 48 x 1 GigE
24 x 1 GigE
2 x 1 GigE
Intel C2D E6600
8 GiB DDR2
4 x 500 GB SATA2
1.75 TB ZFS raidz

PHYSON – MPI

- 32 ядра @ 2 GHz + 160 ядра @ 2,5 GHz
- 384 GiB RAM
- R_{peak} = 1856 Gflops
  (повече от целия Top500 от 06.1993 г.)
- R_{max} = 1507 Gflops = 81% от R_{peak}
  (Open MPI 1.3 + Intel MKL 10.0)

PHYSON – GPU

- Едно Supermicro GPU шаси
  - 1 бр. Xeon E5645
  - 12 GiB ECC DDR3 RAM
  - 2 бр. nVidia Tesla M2090 (512 ядра, 6 GiB GDDR5)
  - 2 x 500 GB HDD RAID1
- 6 ядра (с HT) @ 2,4 GHz + 1024 ядра @ 1,3 GHz
- R_{peak} = 1389 Gflops

PHYSON – FE

- Челен възел – вход към клъстера
- Единствен достъпен по SSH
- Xen гост виртуална машина върху cn001
- Сериозни потребителски ограничения /etc/security/limits.conf

PHYSON – FS

- / – споделен r/o NFS с инсталация на операционната система
- /tmp и /var/volatile – tmpfs за локални цели
- /opt – споделен r/o NFS за приложни програми
- /home – споделен r/w NFS
- /work – споделен r/w NFS (бъдещ Lustre)
- /disk – директно закачен твърд диск

CRNCHR

- За ужас на продавача:
  - 1 бр. AMD Sempron 140
  - 4 GiB DDR3 RAM
    - 2 бр. Sapphire HD5870 (1600 VLIW5 ядра @ 875 MHz)
- ~ 1,5 хил. лв
- 10,8 DP Gflops (CPU) + 1120 DP Gflops (GPU)
- R_{peak} = 1,131 Tflops

ATI/AMD HD5870

- Ядро ATI Cypress
- 1600 поточни ядра @ 875 MHz (850 MHz реф.)
  - 20 SIMD клъстера от по 16 VLIW5 процесора
  - 2 FP MUL/ADD операции на такт
- 2800 SP Gflops / 560 DP Gflops
- 423 Gh/s

Ценова ефективност

- Blue Gene/P – 193,90 лв/Gflops
- PHYSON/MPI – 86,21 лв/Gflops
- PHYSON – 52,39 лв/Gflops
- CRNCHR – 1,33 лв/Gflops

Суперкомпютингът някога беше скъп и недостъпен,
но сега, при наличие на добро желание...


Prestigio

- 20 × HD5850 + 8 × HD6950
- R_{peak} = 12,8 DP Tflops = 46% от R_{peak} на BG/P
- 2,03 Gflops/W
- 1,14 лв/Gflops
- Източник: “Интернет”

“За бога, братя, не купувайте [Blue Gene]!”

Благодарности

- проф. дфзн Ана Пройкова
- HPRI-CT-1999-00026 (TRACS @ EPCC)
- ВУ-Ф 205/2006 (ACL)
- ДО 02-136/2008 (IRC-CoSiM)
- гл.ас. д-р Стоян Писов
- Боян Кроснов

За контакти

- http://physon.phys.uni-sofia.bg
- http://icaci.info
- hristo <при> icaci.info
- xmpp:icaci@jabber.org

Благодаря ви за вниманието!

<3