КОМПЮТЪРНИ КЛЪСТЕРИ

ПРОФ. БОРОВСКА

- *Класификация на клъстерите – според 4 ортогонални атрибута*

1. Конструктивно оформление (packaging):

- ✓ *Компактен клъстер* – възлите са разположени в 1 или повече шкафове в стая, възлите не са свързани към периферия, нар. работни станции без глава, използва системна комуникационна мрежа с висока пропускателна способност и ниска латентност
- ✓ *Хлабав (slack) клъстер* – възлите са свързани към периферни устройства т.е. те са пълноценни SMP, работни станции или PC, могат да бъдат разположени в различни стаи, сгради, географски отдалечени райони

2. Управление

- Централизирано– всички възли са притежавани контролирани , управлявани & администрирани от централен оператор (обикновено компактен клъстер)
- Децентрализирано – възлите имат индивидуални собственици ; собственикът може да реконфигурира, upgrade или даже да изключи работната станция, когато пожелае

Хлабавият клъстер може да бъде контролиран или управляван както централизирано, така и децентрализирано.

3. Хомогенност

- Хомогенен клъстер – всички възли имат една и съща платформа
- Хетерогенен клъстер – възлите имат различни платформи, миграцията на процеси не е възможна

4. Сигурност

- ✓ *"Отворена" интраклъстерна комуникация* – лесна за имплементиране, но външна машина може да осъществи достъп до комуникационните пътища и така и до индивидуалните възли, използвайки стандартни протоколи (т.е., TCP/IP)

Недостатъци:

- *Интраклъстерната комуникация не е сигурна*
- *Външните комуникации могат да нарушат интраклъстерните комуникации по непредвидим начин*
- *Стандартните комуникационни протоколи имат високи допълнителни разходи*

- ✓ *"Затворена" интраклъстерна комуникация* – изолирана от външния свят, недостатък - липсата на стандарт за ефективна интраклъстернакомуникация

Специализиран клъстер (Dedicated cluster)

- Типично се инсталира в един шкаф в централна компютърна зала
- *Има хомогенна конфигурация с еднотипови възли*
- *Управлява се от една администраторска група*
- *Достъпът до него се осъществява чрез front-end система*
- Използва се като заместител на високо производителните компютри
- *Инсталира се, използва се и се администрира като една машина*
- Изпълнява както интерактивни, така и batch jobs
- ***Висока пропускателна способност & намалено време за отговор***

Клъстер на организации (Enterprise cluster)

- Използват се главно с цел ***използване на свободните ресурси във възлите***
- *Всеки възел обикновено е SMP, работна станция, или PC, със свързана периферия*
- *Възлите индивидуално се притежават от множество собственици*; локалните програми на собственика имат *по-висок приоритет* от програмите на организацията
- Типично *възлите са географски разпределени*
- Конфигуриран е с хетерогенни компютърни възли, свързани чрез евтина мрежа Ethernet

АСПЕКТИ

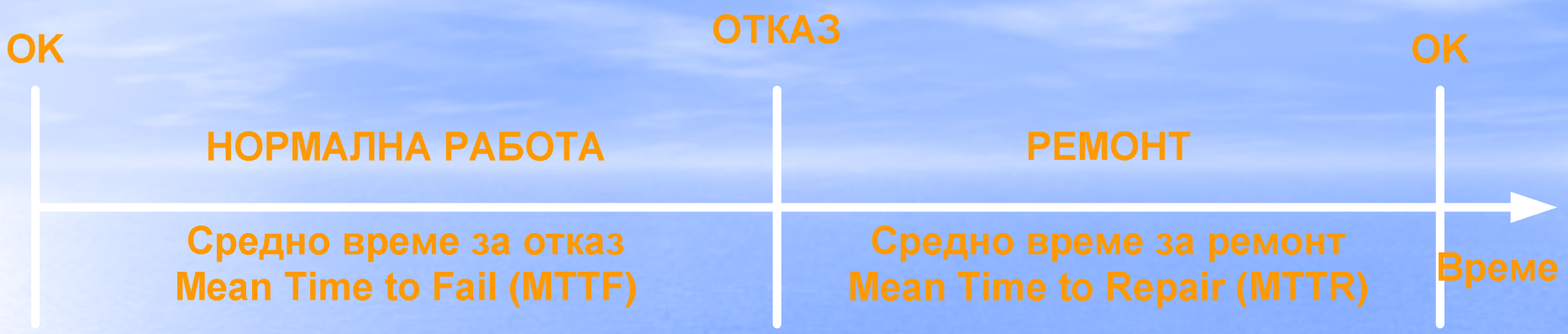
- *Надеждност* – множество процесори, памети, дискове, В/И устройства, мрежи, и т.н.
- *Единна система* – чрез клъстерирането на множество работни станции се получава единна система, еквивалентна на една огромна работна станция, наречена *мегастанция*
- *Управление на задачите – пакетна обработка, баланс на товара, паралелна обработка*
- *Ефективна комуникация – често се използват конвенциални мрежи (Ethernet, ATM) със стандартни комуникационни протоколи (високи доп. разходи), дългите връзки обуславят по-голяма латентност & смущения*

НАДЕЖДНОСТ, ДОСТЪПНОСТ, ЛЕСНОТА НА ОБСЛУЖВАНЕТО

RAS – RELIABILITY, AVAILABILITY, SERVICEABILITY

- ***Надеждност*** – показва времето, през което системата функционира без отказ
- ***Достъпност*** – показва процента от времето, през което потребителят може да използва системата
- ***Лесnота при обслужването*** – включително поддръжката на хардуера & софтуера, ремонта, upgrade, и т.н..

ДОСТЪПНОСТ

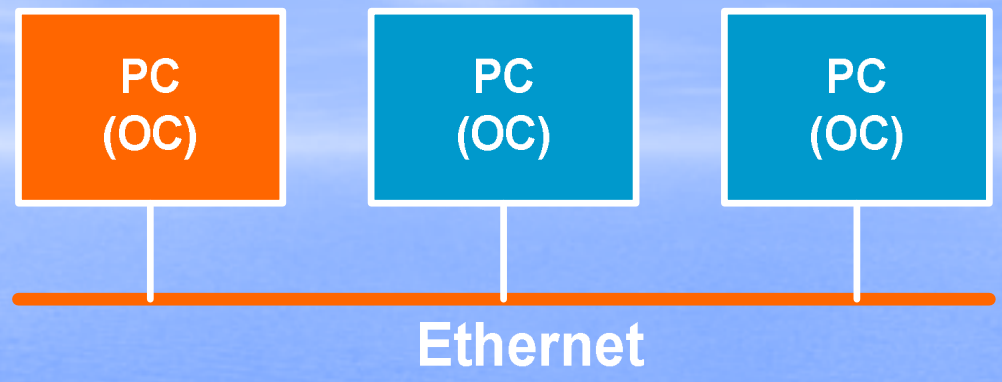


Цикълът работа - ремонт на компютърна система

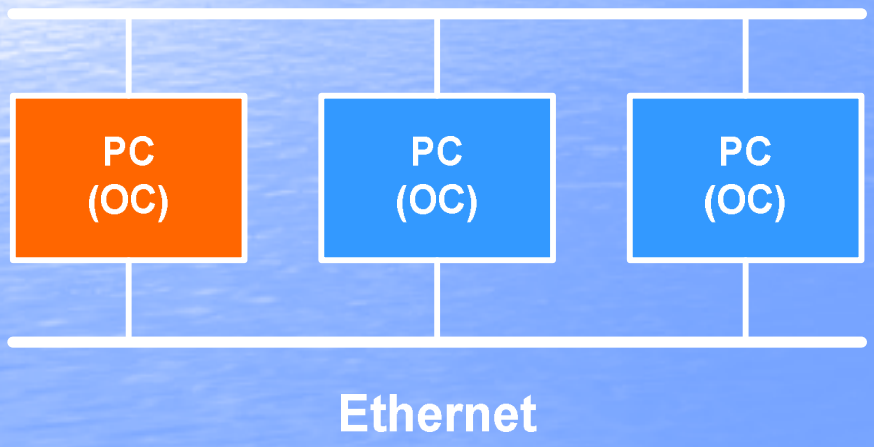
- Надеждността на системата се измерва чрез среното време за отказ, което представлява средното време на работа преди да възникне отказ в системата (или нейн компонент)
- Леснотата на обслужване се определя от средното време за ремонт на системата MTTR
- **Надеждност = MTTF/(MTTF+MTTR)**

Единични точки на отказ в клъстерите
Single points of failure in clusters

– хардуерни или софтуерни компоненти, чийто отказ води до отказ на цялата система



КЛЪСТЕР ОТ РАБОТНИ СТАНЦИИ



Клъстер с две мрежи

Клъстер с общ диск

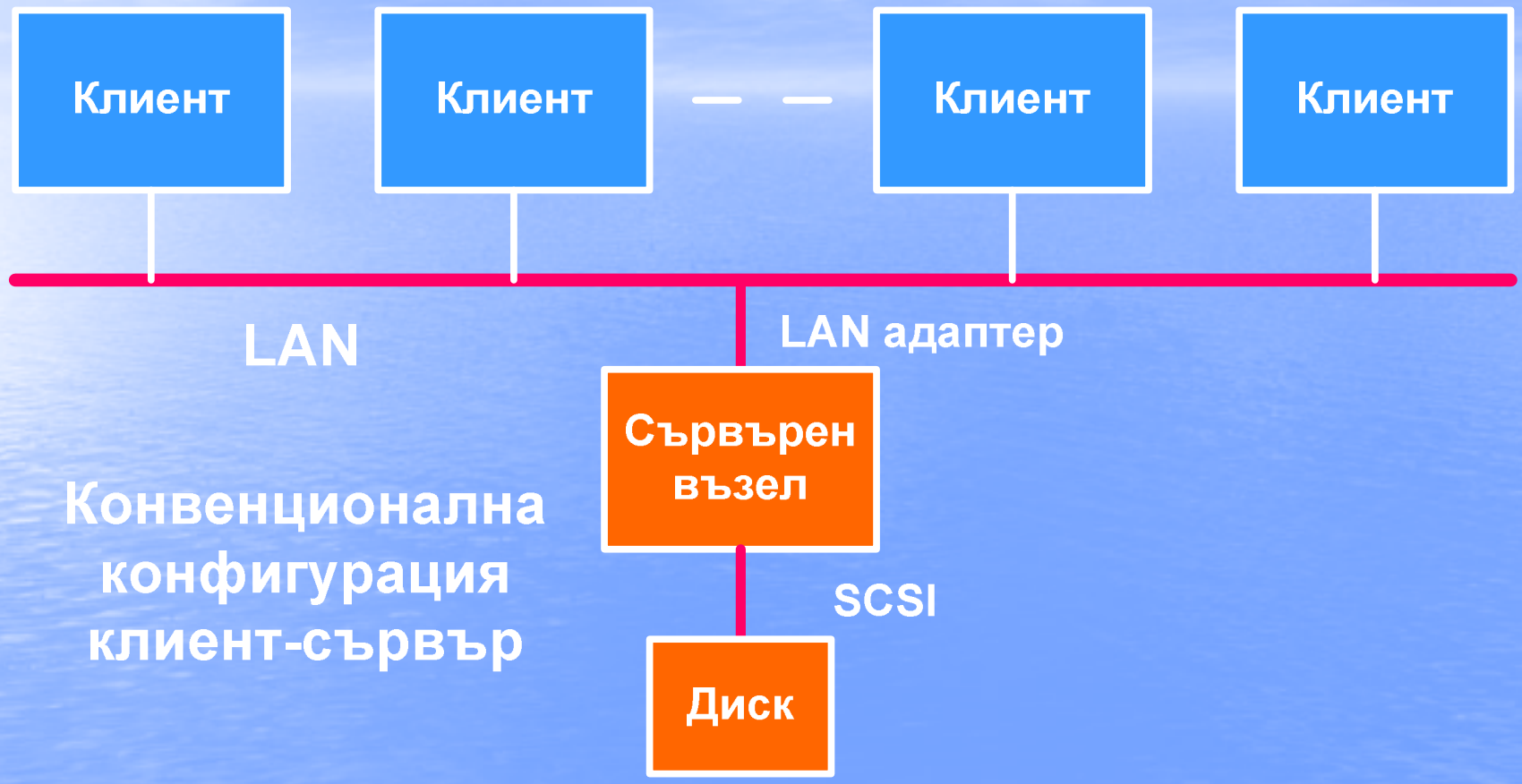
Техники за повишаване на надеждността

- *Увеличаване на MTTF*
- *Намаляване на MTTR*
- *Изолиране на компонентите* – когато *основният* компонент откаже, обслужването се поема от *backup* компонент; основният & backup компоненти са *изолирани един от друг*, така че не са подложени на въздействието на едни и същи причини за отказ;

Елиминират се единичните точки на отказ

Отказалият компонент се ремонтира докато останалата система работи

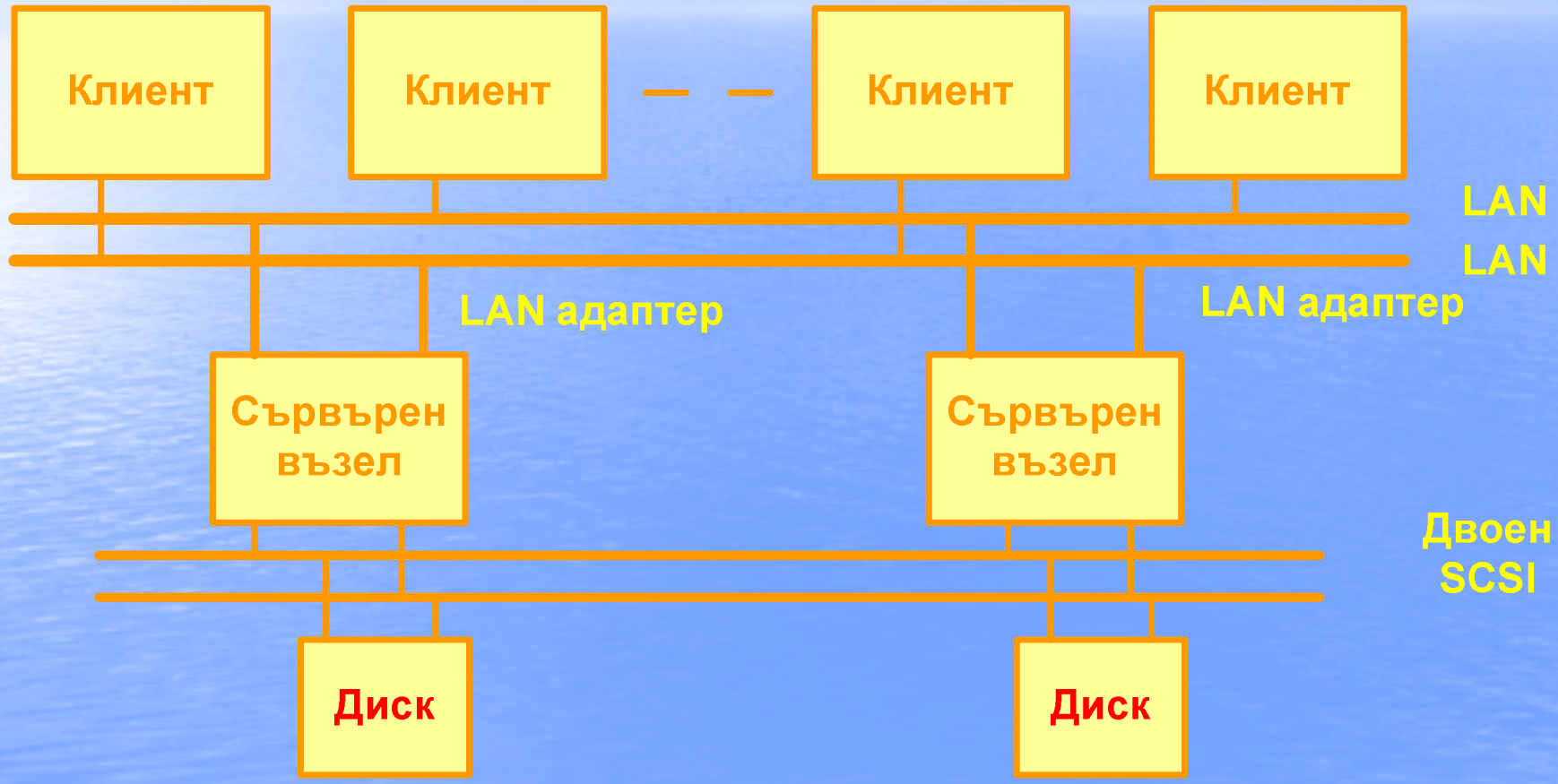
Единични точки на отказ в клъстерите



Конвенционална конфигурация клиент-сървър

5 възможни single points of failure:
(1)LAN; (2)LAN адаптера на сървърния възел,
(3)сървър; (4)SCSI; (5)външен диск

Дублиране на ресурсите на клъстера за елиминиране на всички single points of failure

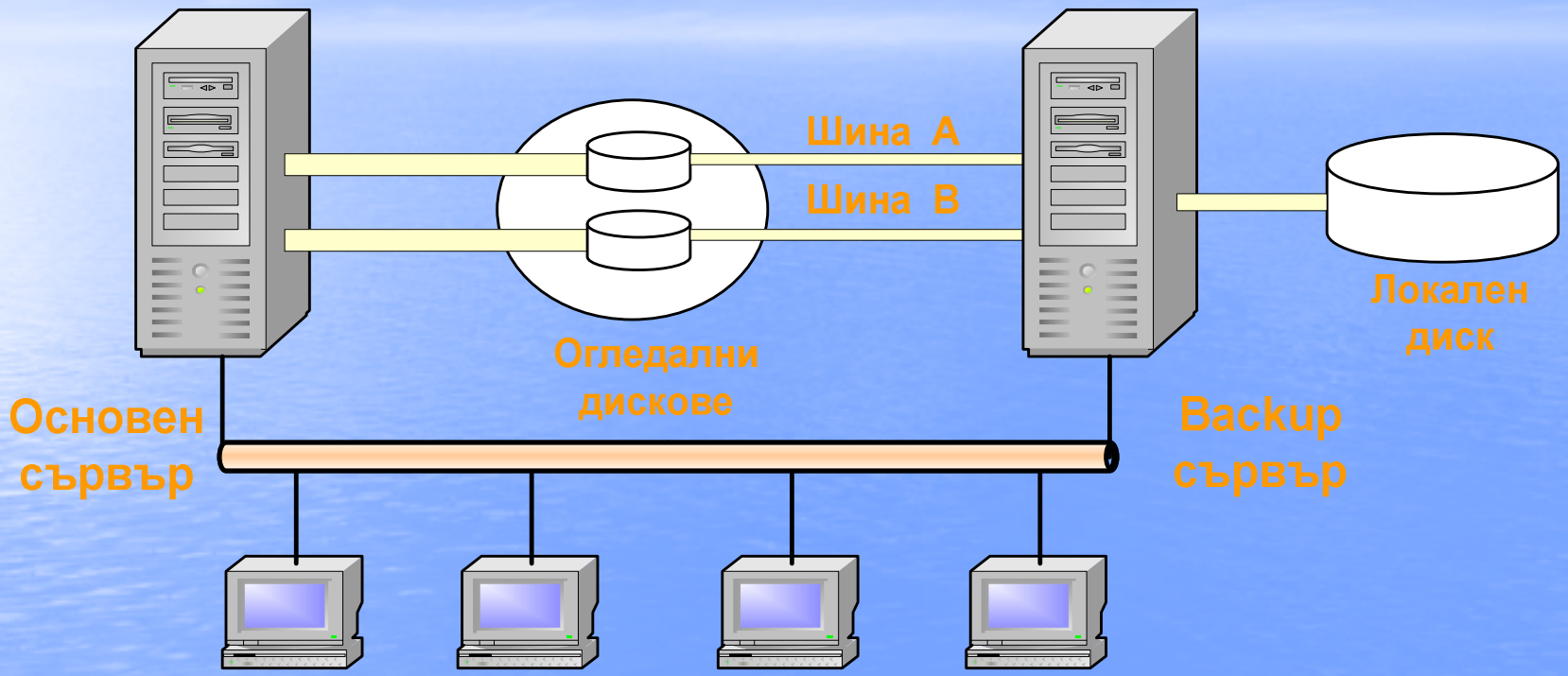


Система с висока надеждност
(с агресивен излишък)

Конфигурации на дублиращите компоненти

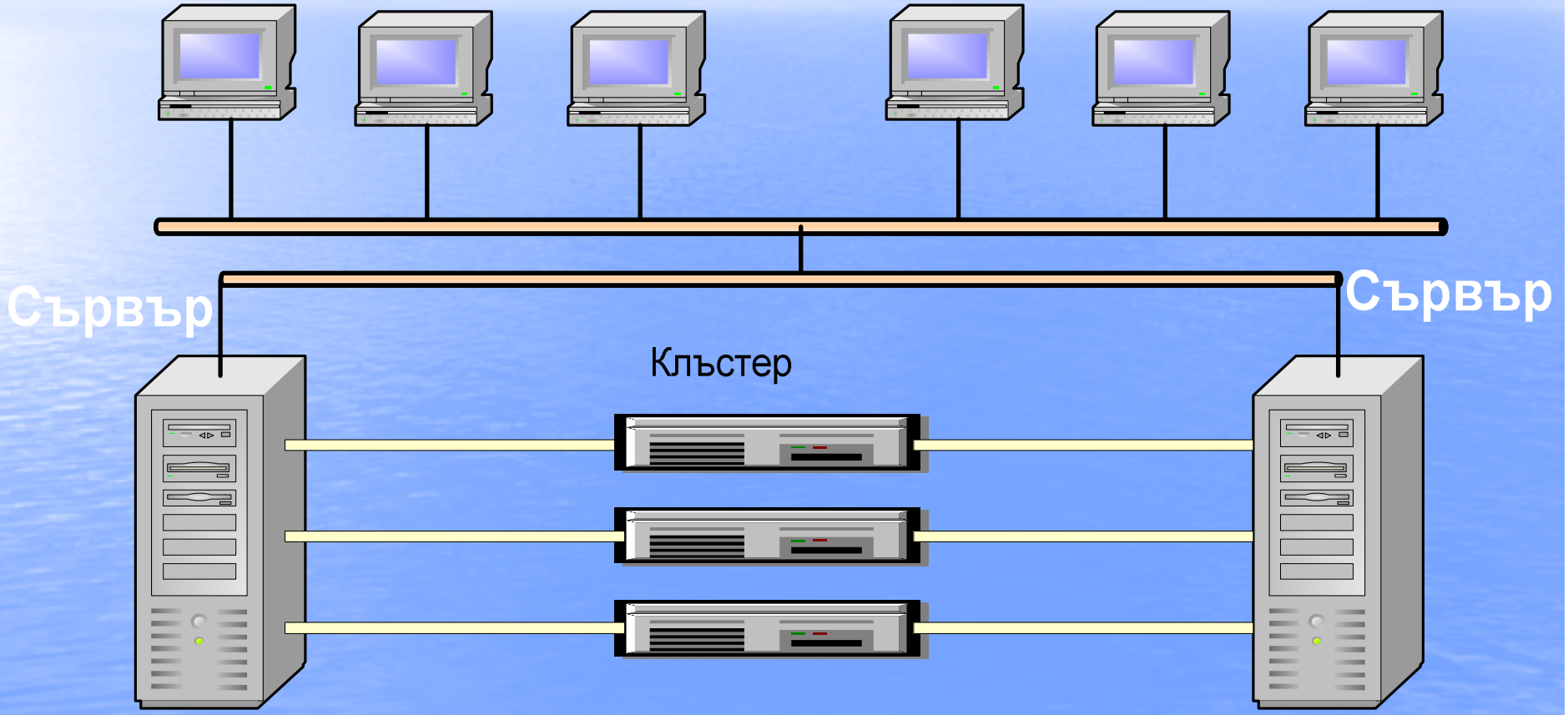
- ***Hot Standby*** – основният компонент работи, backup компонентът е готов (hot) да поеме работата в случай на отказ на основния (икономично решение – 1 standby компонент да поддържа (back up) множество основни компоненти)
- ***Mutual Takeover*** – всички компоненти са основни; 1 компонент откаже – работният му товар се разпределя между останалите изправни компоненти
- ***Fault-tolerant*** – N компонента осигуряват производителността на само 1 компонент

Hot Standby Multiserver Clusters



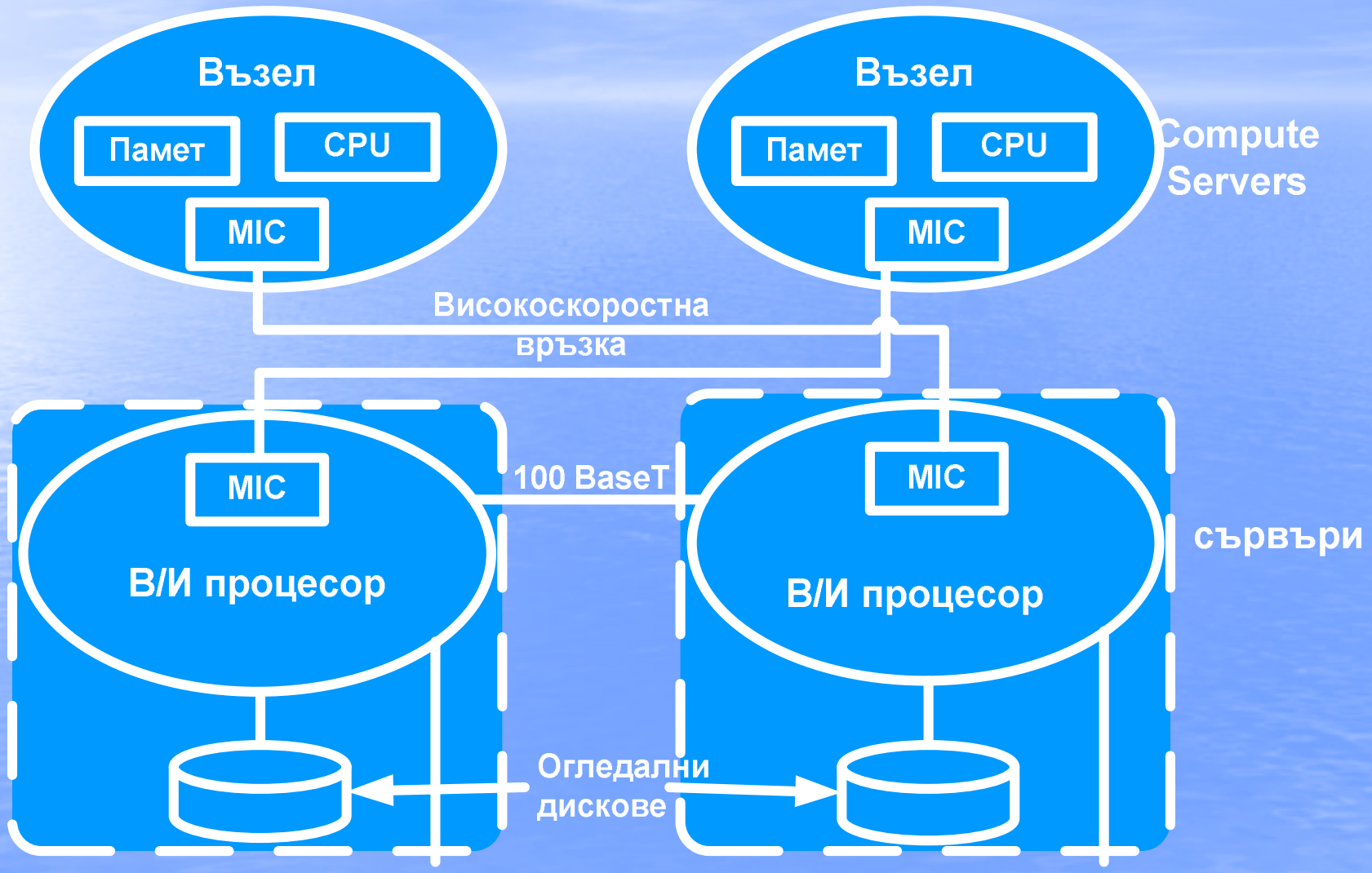
The hot standby cluster configuration

Архитектура на клъстер с два активни възела



- ***Failover*** – *най-важното качество, което се изисква при съвременните клъстери за комерсиални приложения*; при отказ на даден компонент, останалата част поема функциите на отказалия компонент; осигурява диагностика на отказа, издава съобщение за отказа & възстановяване
- *Диагноза на отказа при клъстери с дуални мрежи:* всеки възел има *heartbeat daemon,* който периодично изпраща heartbeat message към *master node* през двете мрежи
  - ✓Отказът е във възела
  - ✓Отказът е при връзка към мрежата

Fault-Tolerant Multiserver Cluster



КЛЪСТЕРНИ ПРОДУКТИ

- Сега *в света се използват* повече от *100 000 компютърни клъстери*
- Те включват както *комерсиални клъстери, така и custom-designed clusters*
- *В повечете случаи възлите са PCs, работни станции & SMP сървъри*
- Размерът на клъстерите в повечето случаи е от порядъка на десетки възела; малко клъстери имат повече от 100 възела
- Повечето клъстери използват *commodity networks* като Fast или Gigabit Ethernet, FDDI rings, ATM or Myrinet switches освен регулярните LAN връзки между възлите

Клъстери от SMP сървъри

Очевидна индустриална тенденция е да се клъстерират определен брой хомогенни SMP сървъри в рамките на интегриран *superserver*

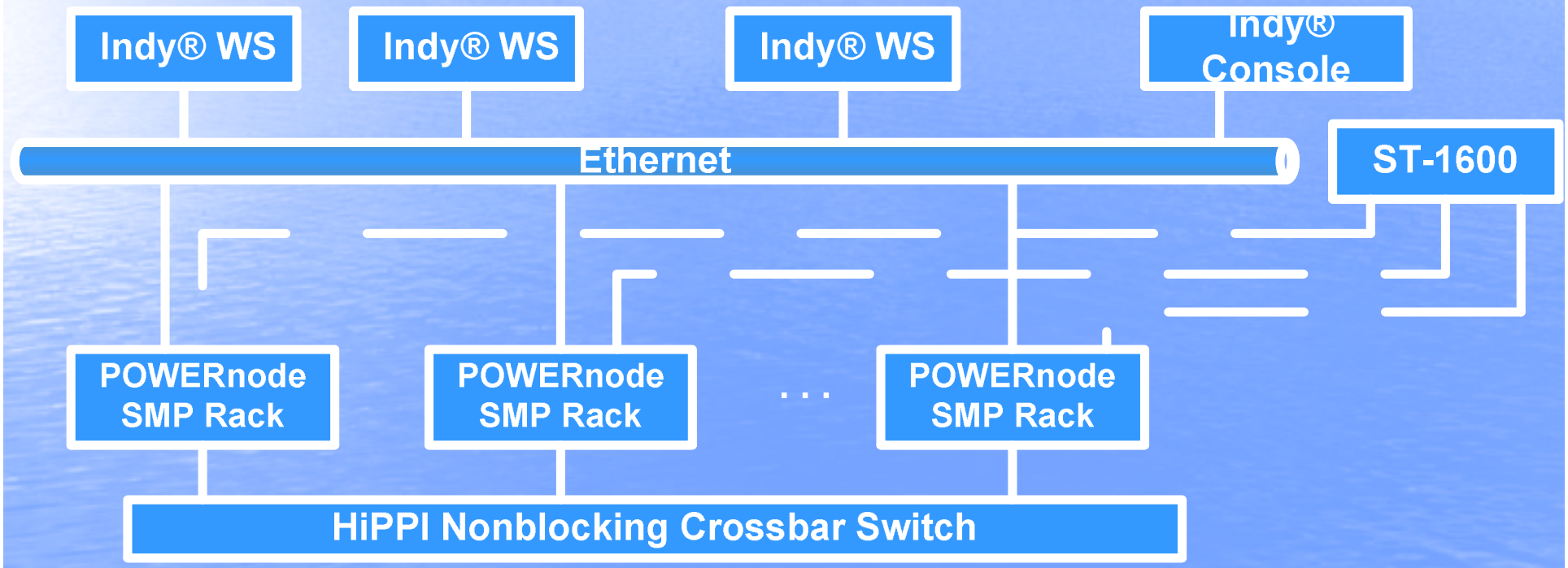
SGI POWER CHALLENGEarray

- Системата свързва *2-8 SMP възела*, наричани POWERвъзли, формиращи клъстера като

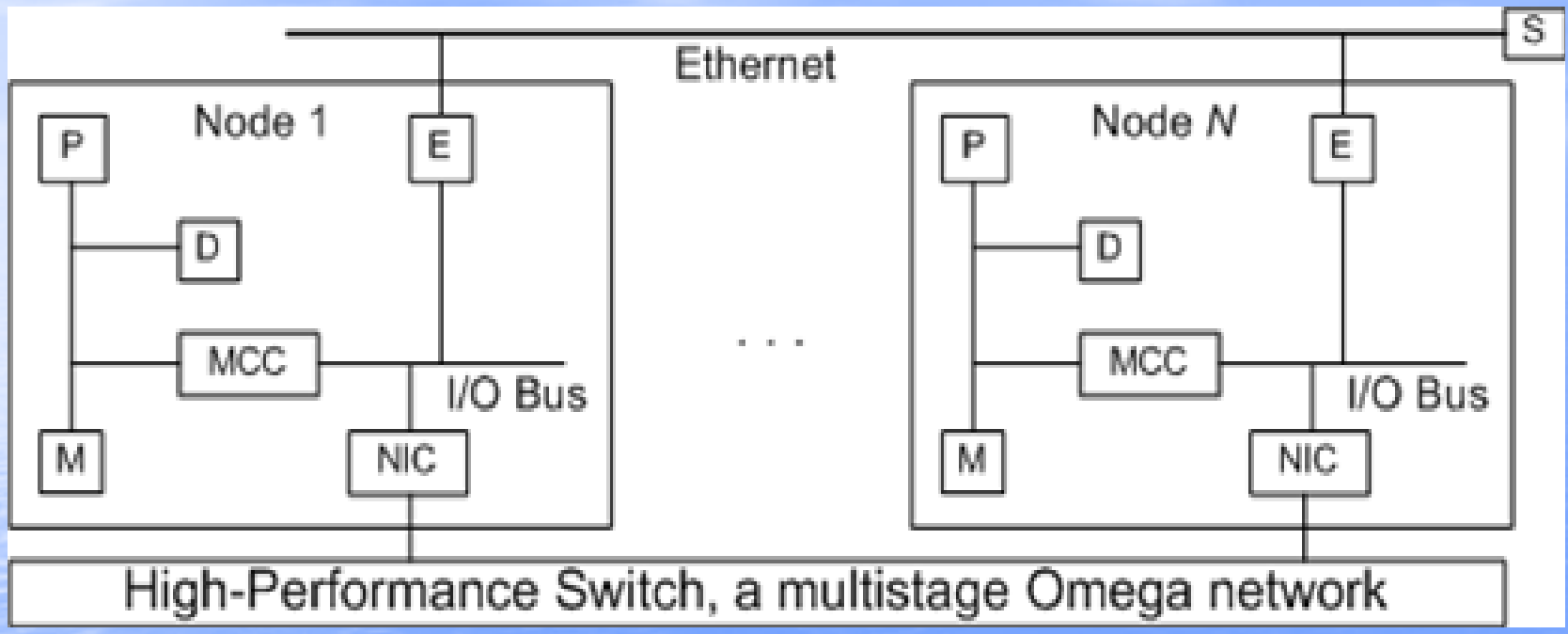
superserver

- Всеки POWERвъзел е Power Challenge SMP server, обхващащ 36 MIPS R10000 processors & 16 GB обща памет
- Общо, суперсъвърният клъстер може да осигури до 128 GB главна памет, повече от 4 GB/s скорост на трансфер от диска
- Възлите са свързани чрез crossbar HiPPI switch за осигуряване на високоскоростна комуникация
- Достъпът до системата може да се осъществи чрез Ethernet и работни станции Indy

Клъстер от SMP сървъри на Silicon Graphics

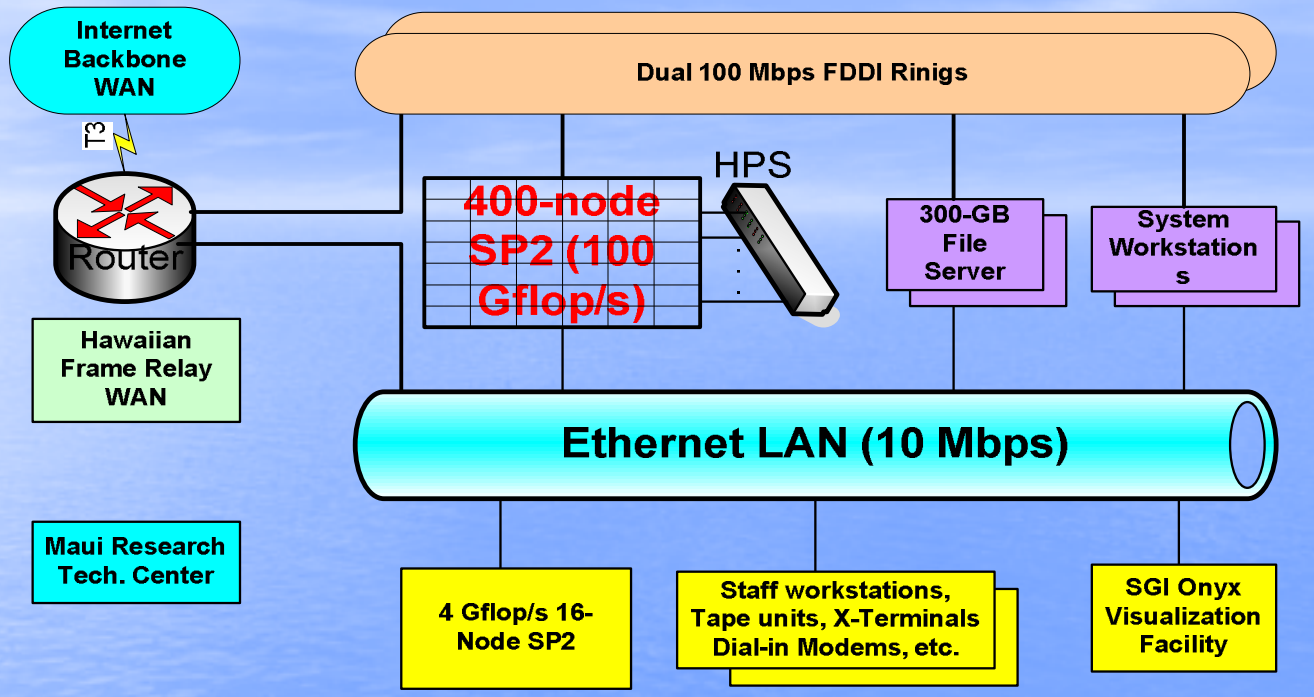


Системна архитектура на IBM SP2



P: processor, M: memory, D: disk
MCC: MicroChannel controller
NIC: network interface switch
E: Ethernet adapter
S: system console

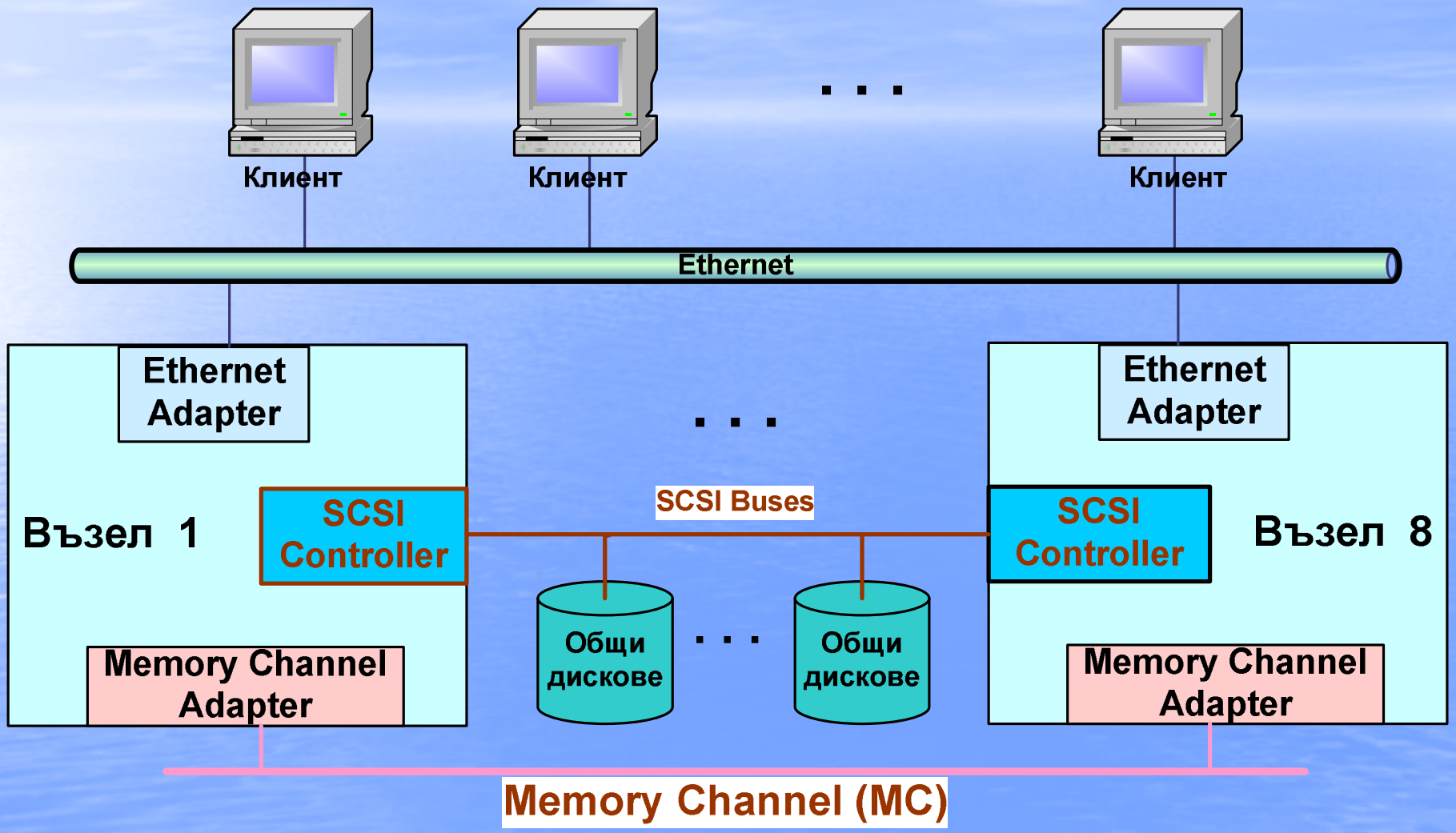
Мултикомпютър SP2 с 400 възела в Maui High Performance Computing Center (Hawaii)



Legends

Symbol	Description
	Ethernet
	FDDI = Fiber Distributed Data Interface
	High-performance Switch
	T3 line (45 Mbps), and T1 lines (1.54 Mbps). Thin links are 10-Mbps Ethernet connections
	Router

Digital TruCluster



IBM Cluster

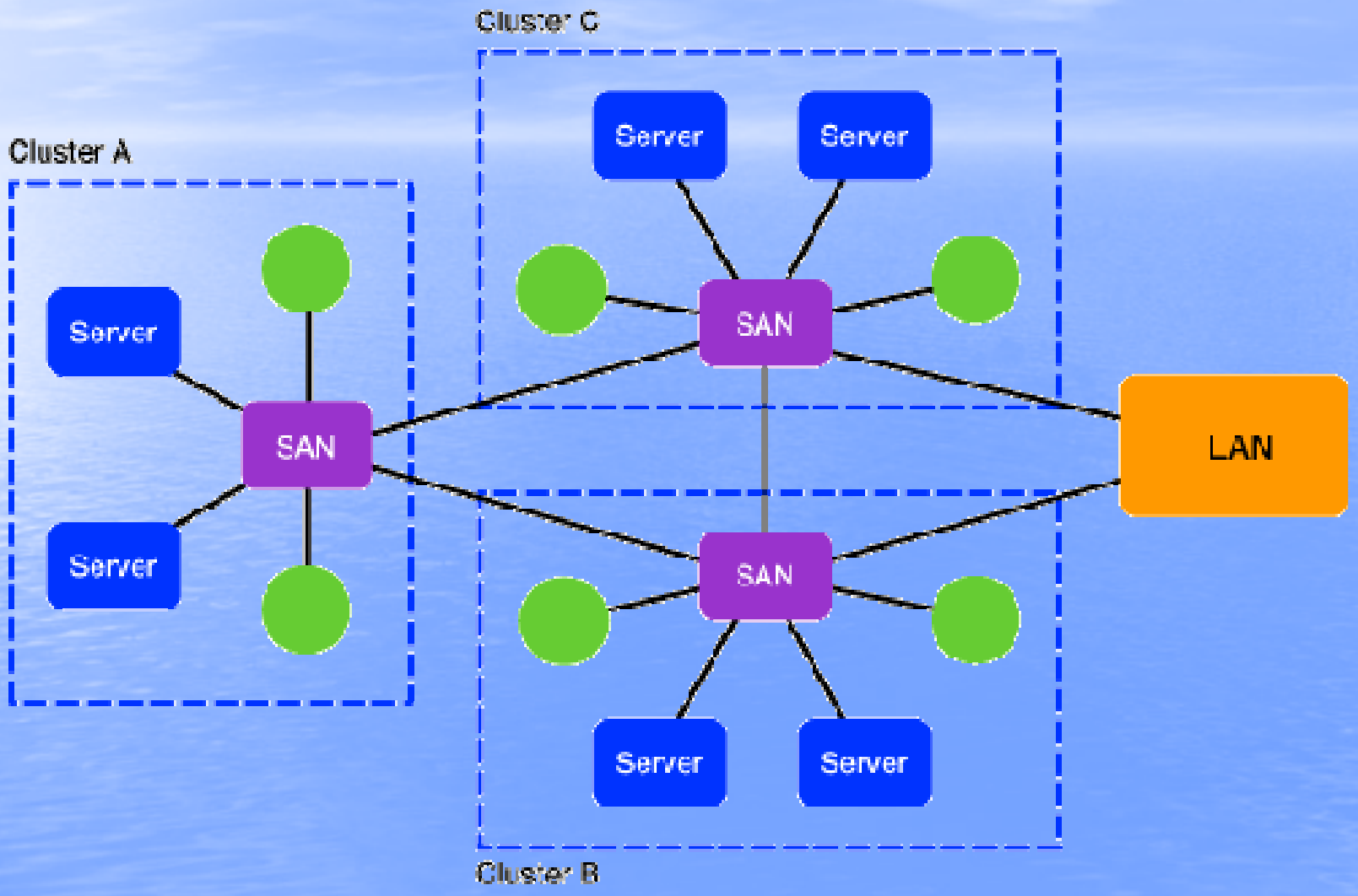


Figure 5.

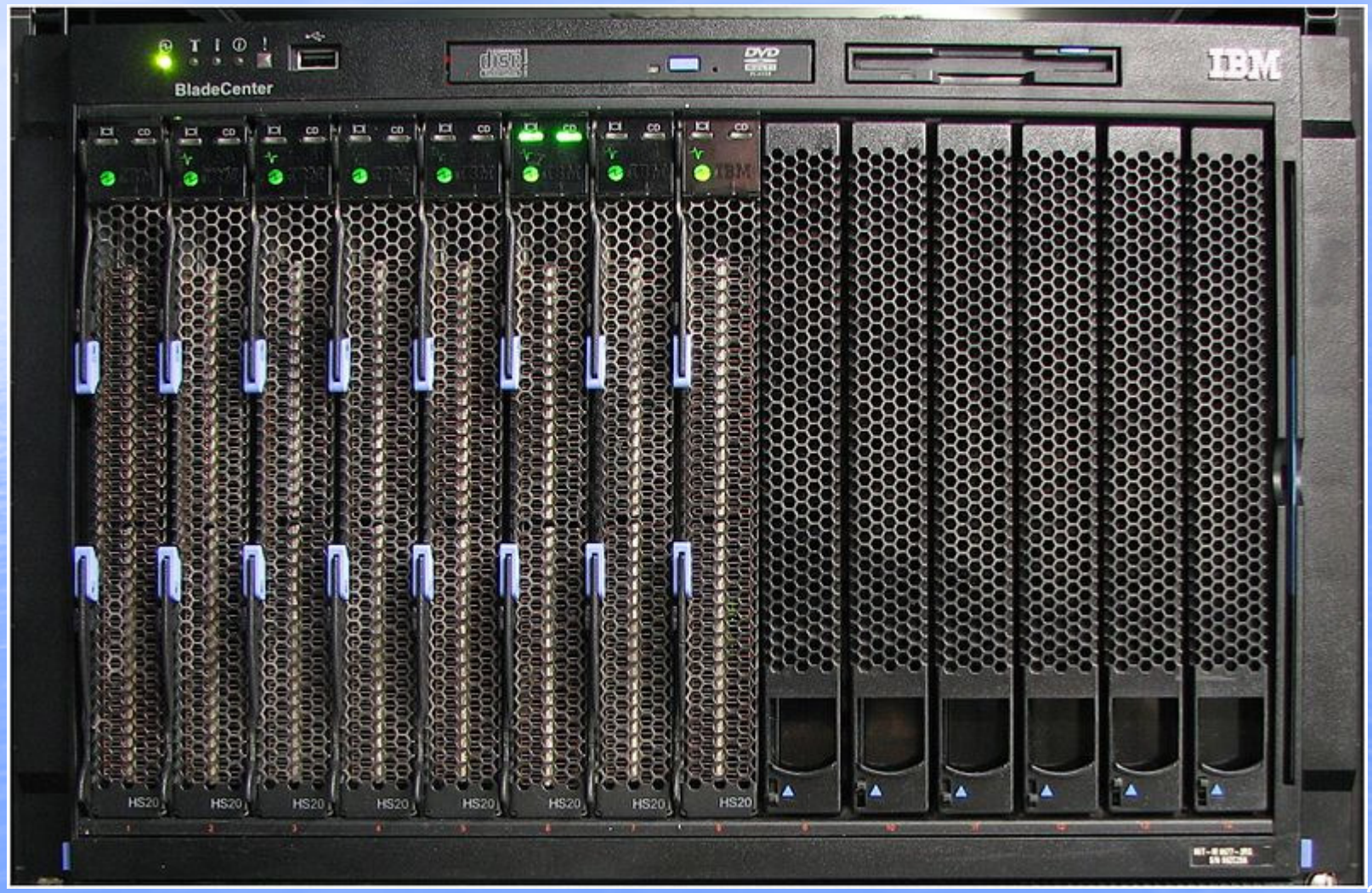
Pleiades

Pleiades Site
NASA/Ames Research Center/NAS System Family SGI Altix System Model SGI Altix ICE 8200

Carnegie Mellon University – Beowulf Distributed Computer Cluster

16 node Linux cluster interconnected with Myrinet 2000

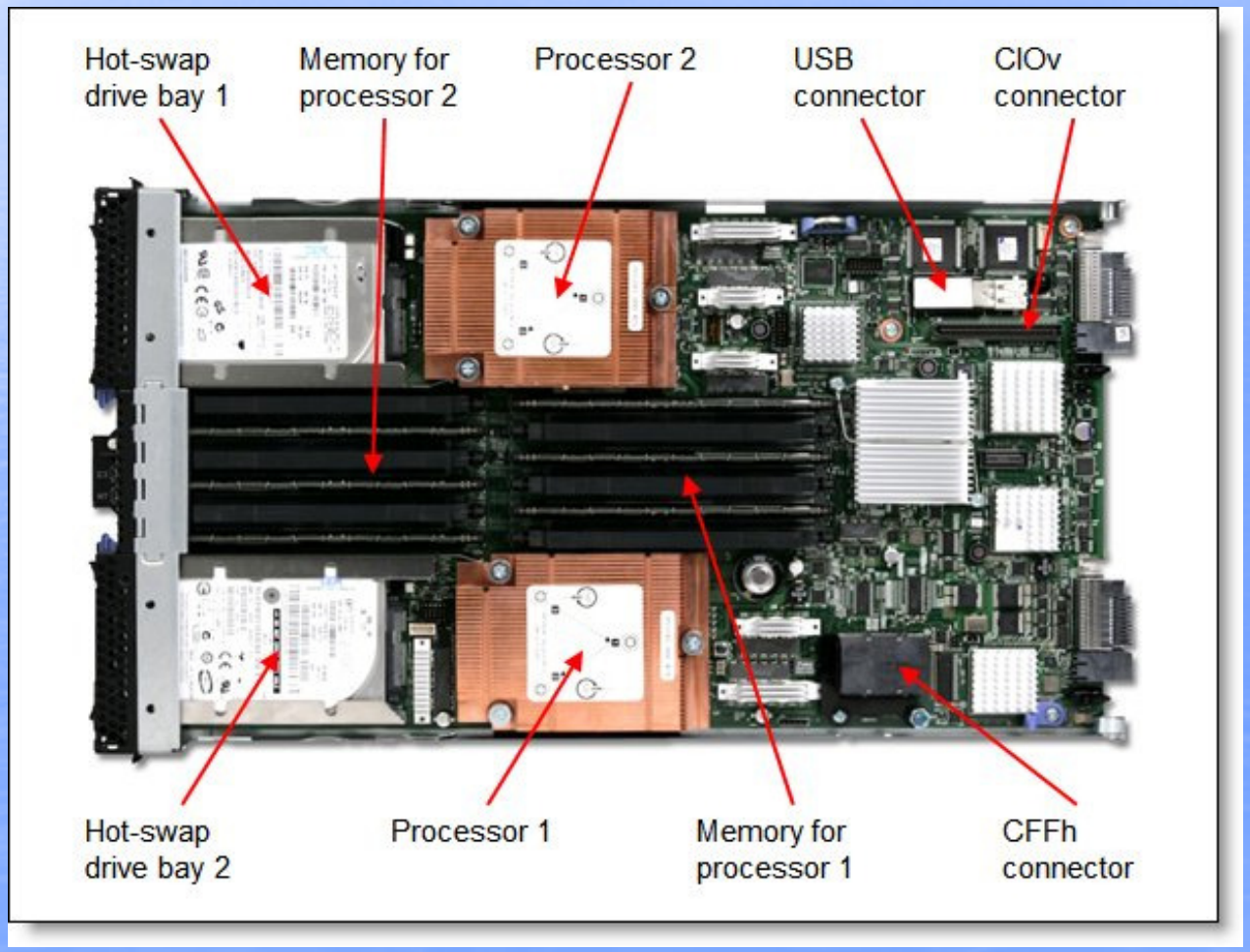
IBM BladeCenter E front side: 8 blade servers (HS20) followed by 6 empty slots



BladeCenter E back side

Magerit supercomputer (CeSViMa) has 86 Blade Centers (6 Blade Center E on each computing rack)

The IBM® BladeCenter® HS22 is a two-socket blade server running Intel Xeon processors.



IBM® BladeCenter® HS22

- Intel Xeon 5600 series processors, up to 3.6 GHz
- Up to 12 MB L3 cache
- Models with Intel Xeon 5600 series processors: 192 GB
- OS - Microsoft® Windows®, Red Hat Enterprise Linux®, SUSE Linux Enterprise, VMware, Oracle Solaris
- Remote management - IBM Integrated Management Module (IMM)


In-silico Drug Design


IBM
IBM

Sun constellations linux cluster

http://www.youtube.com/watch?v=gmGlrpjsauM&feature=related

FreeStudio.exe

http://www.dvdvideosoft.com/products/dvd/Free-YouTube-Download.htm

КРАЙ