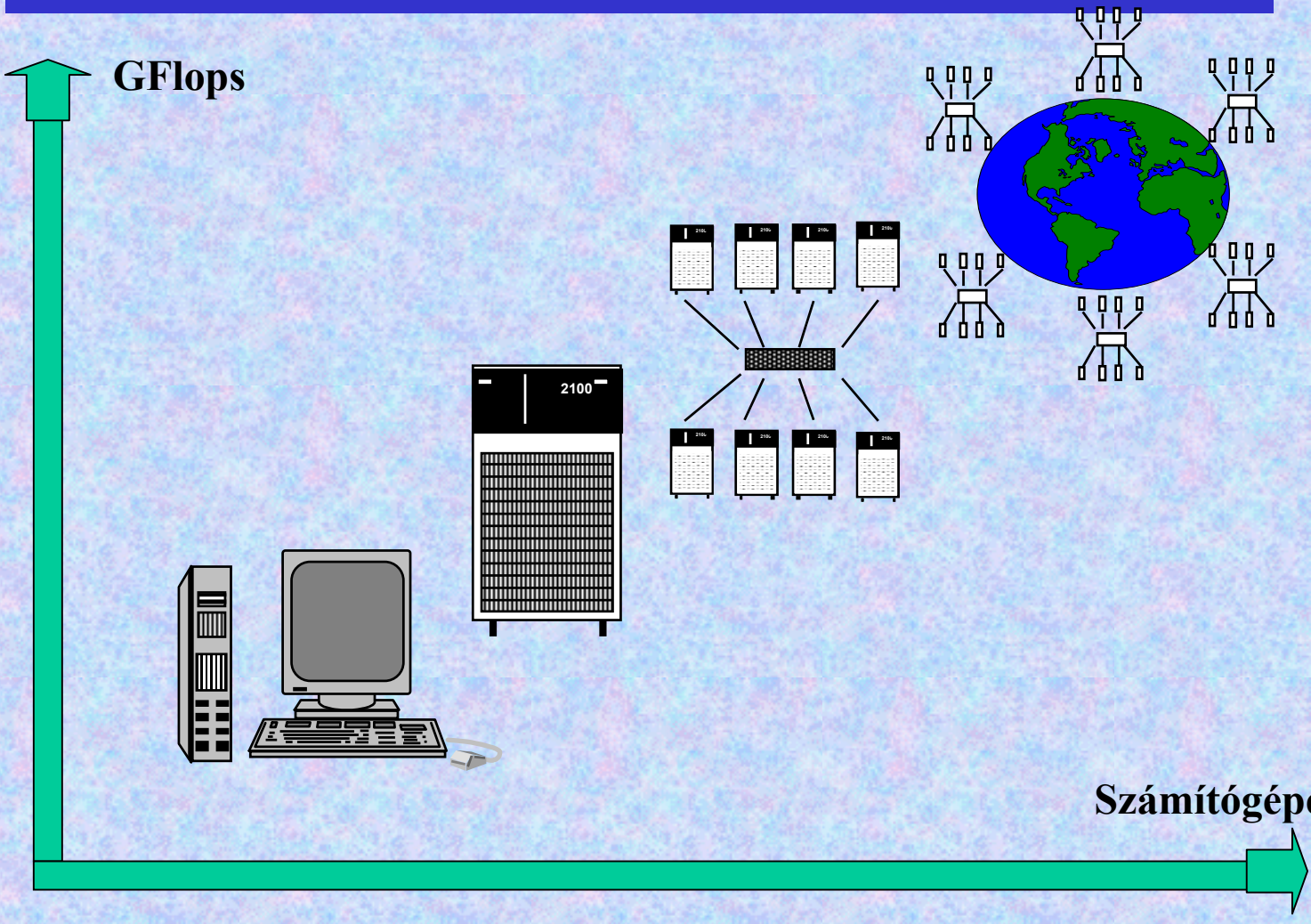
Szuperszámítógépektől a klaszterekig

Kacsuk Péter
www.lpds.sztaki.hu

Tartalomjegyzék

Nagysebességű rendszerek fejlődése



A szuperszámítógépek megalkotásának eredeti motivációi

- Az un. nagy kihívást jelentő problémák megoldása heteket sőt hónapokat vett igénybe még a mainframe számítógépeken is

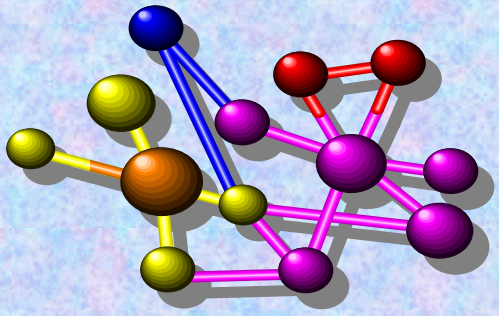
- Sok processzort kellett összekapcsolni speciális nagysebességű belső hálózatokkal annak érdekében, hogy a fenti problémákat ésszerű időn belül meg lehessen oldani

Sequential Architecture Limitations

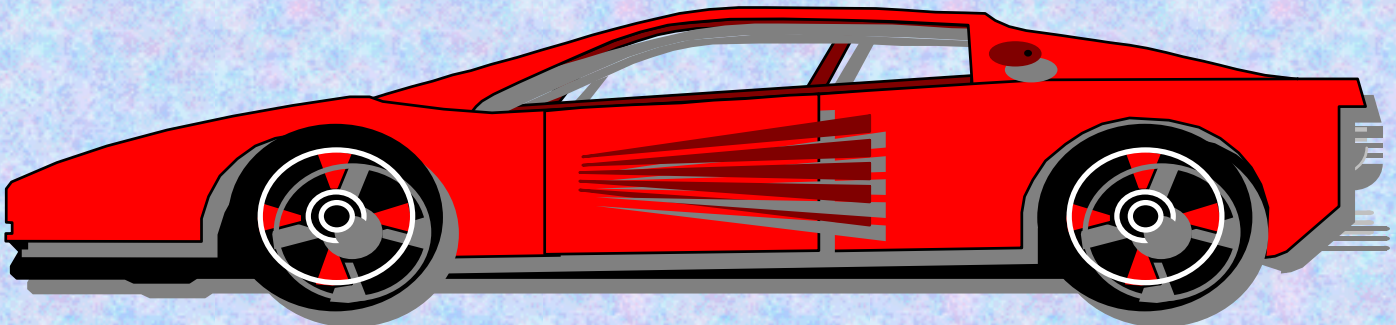
- Sequential architectures reaching physica limitation (speed of light, thermodynamics)
- Hardware improvements like pipelining, Superscalar, etc., are non-scalable and requires sophisticated Compiler Technology.
- Vector Processing works well for certain kind of problems.

Alkalmazási területek

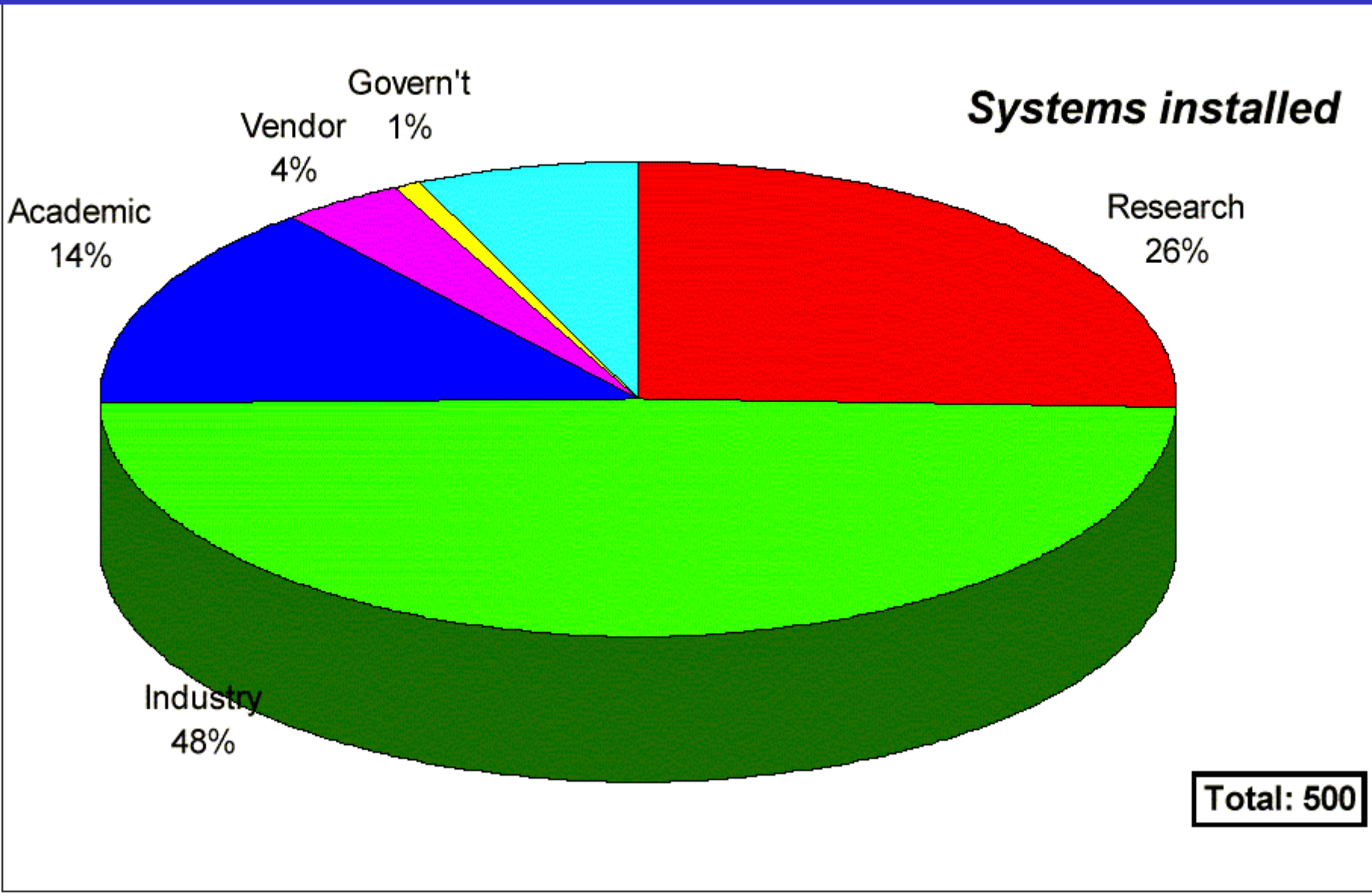
- **Tudományos számítások**
  - **csillagászati modellezések**
  - **molekuláris biológia**
  - **vegyészeti kutatások**
  - **atomfizika, stb.**

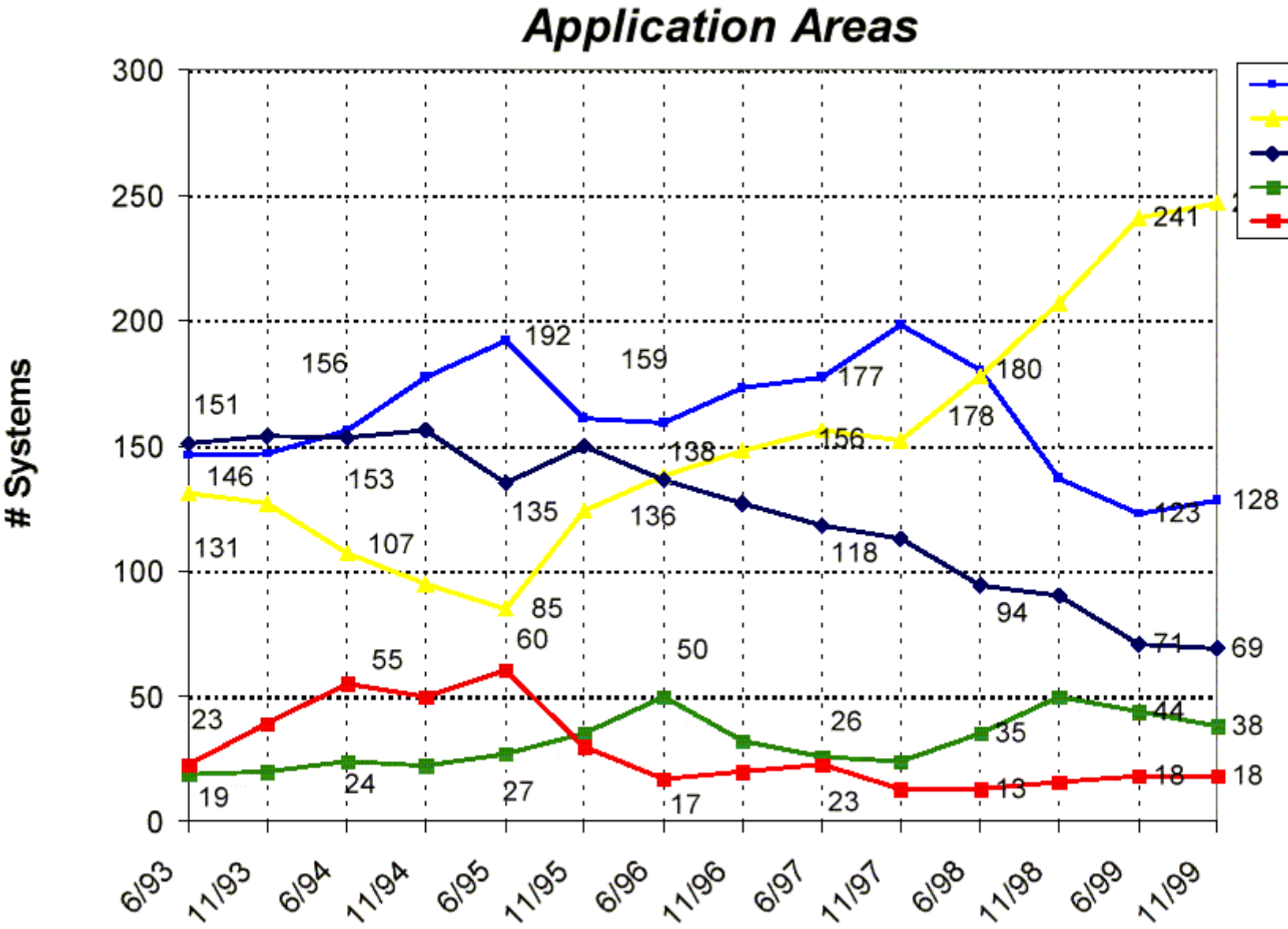


- **Mérnöki számítások**
  - **gépjárműipar (ütközésmodellezés, áramlástani formatervezés, stb.)**



TOP 500 számítógép alkalmazási területeinek megoszlása





A szuperszámítógép korszak

- Áttörés: Transzputerek (1984)
- Fő típusok:
  - Vektorprocesszorok
    - Cray, NEC, Hitachi
    - túl drága
  - Szimmetrikus multiprocesszorok
    - Sequent, Sun, SGI, IBM
    - korlátozott mértékben bővíthető
  - MPP (Masszívan párhuzamos processzorok)
    - Cray T3E, Intel Paragon, SGI Origin-2000

Taxonomy of Architectures

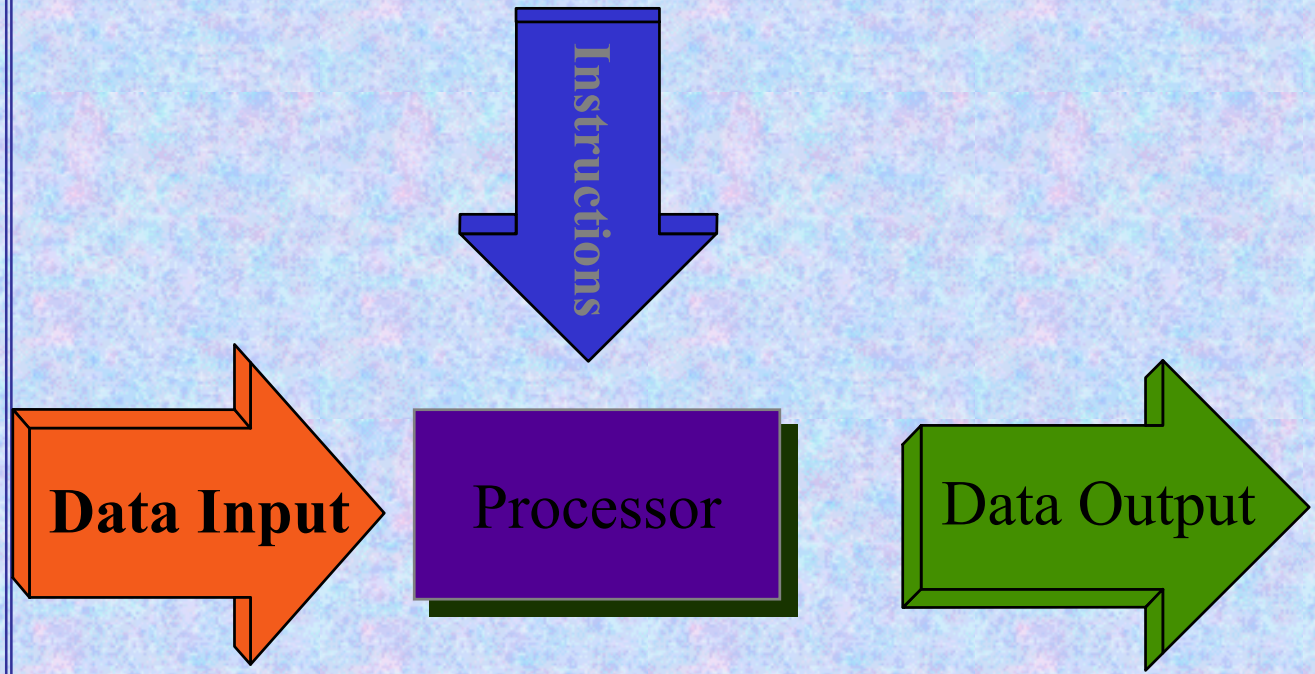
- Simple classification by Flynn:

  (No. of instruction and data streams)

  - **SISD - conventional**
  - **SIMD - data parallel, vector computing**
  - **MISD - systolic arrays**
  - **MIMD - very general, multiple approaches.**

- Current focus is on MIMD model, using general purpose processors or multicomputers.

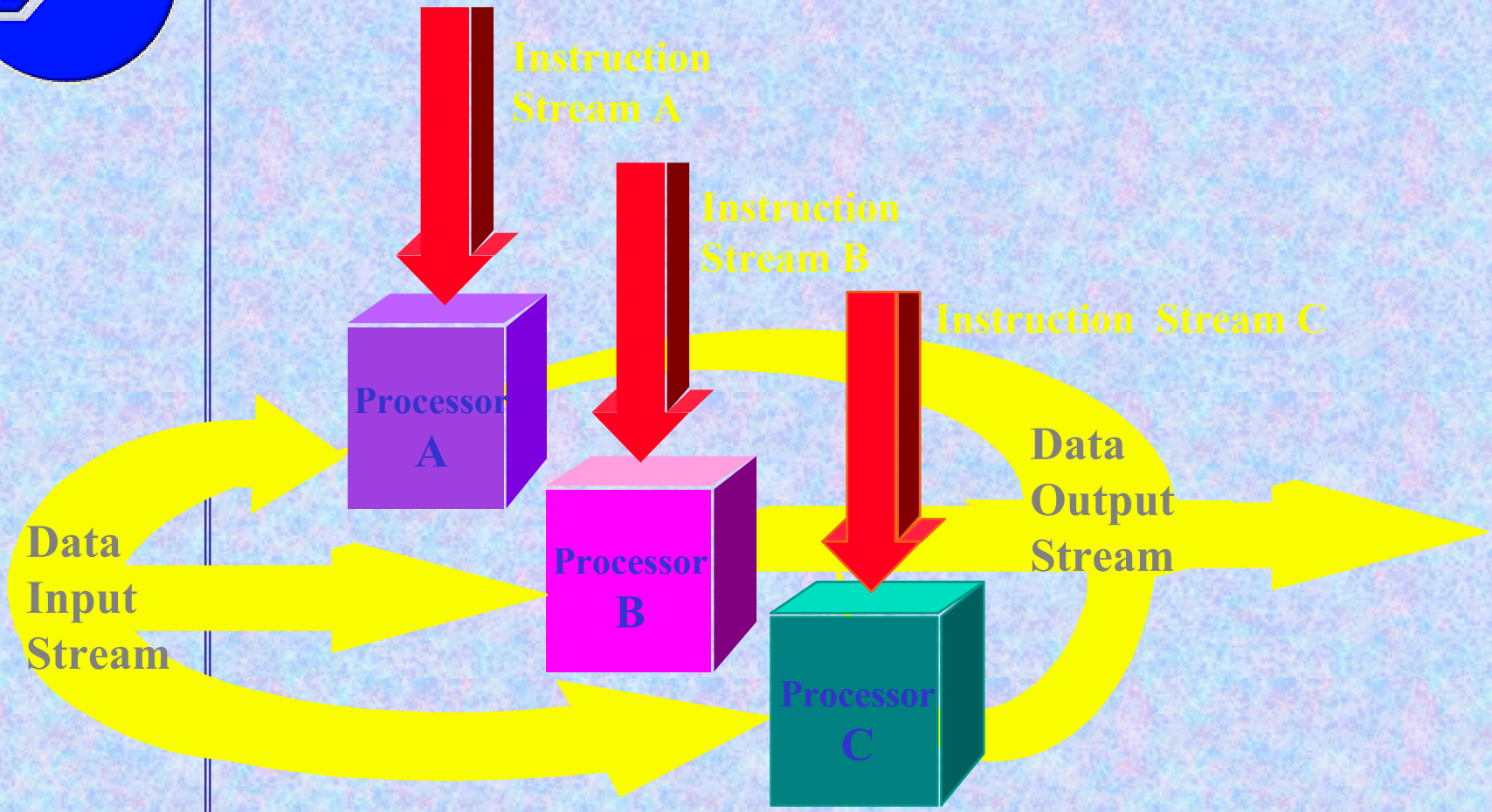
SISD : A Conventional Computer



➔Speed is limited by the rate at which computer can transfer information internally.

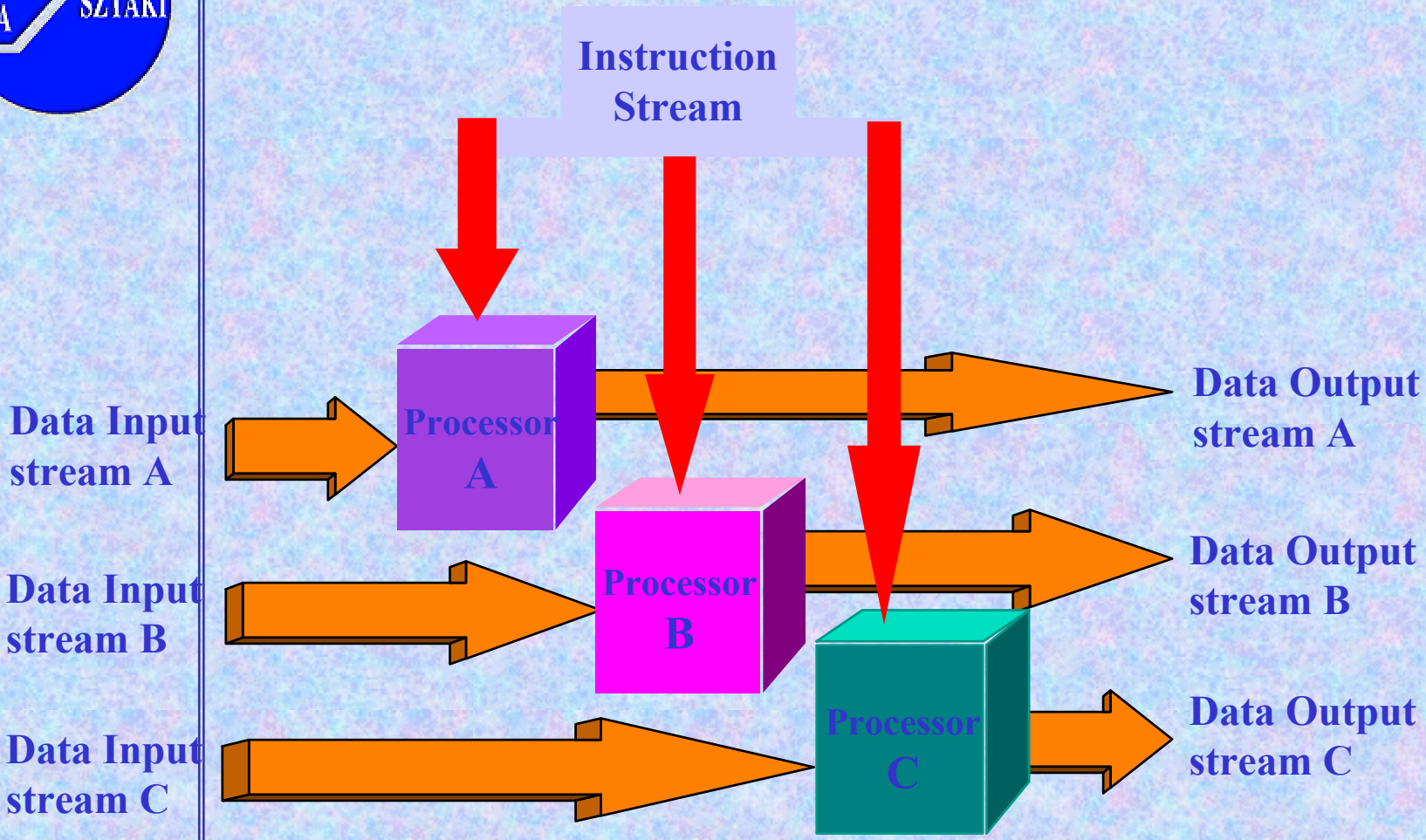
Ex:PC, Macintosh, Workstations

The MISD Architecture



➔More of an intellectual exercise than a practical configuration.
Few built, but commercially not available

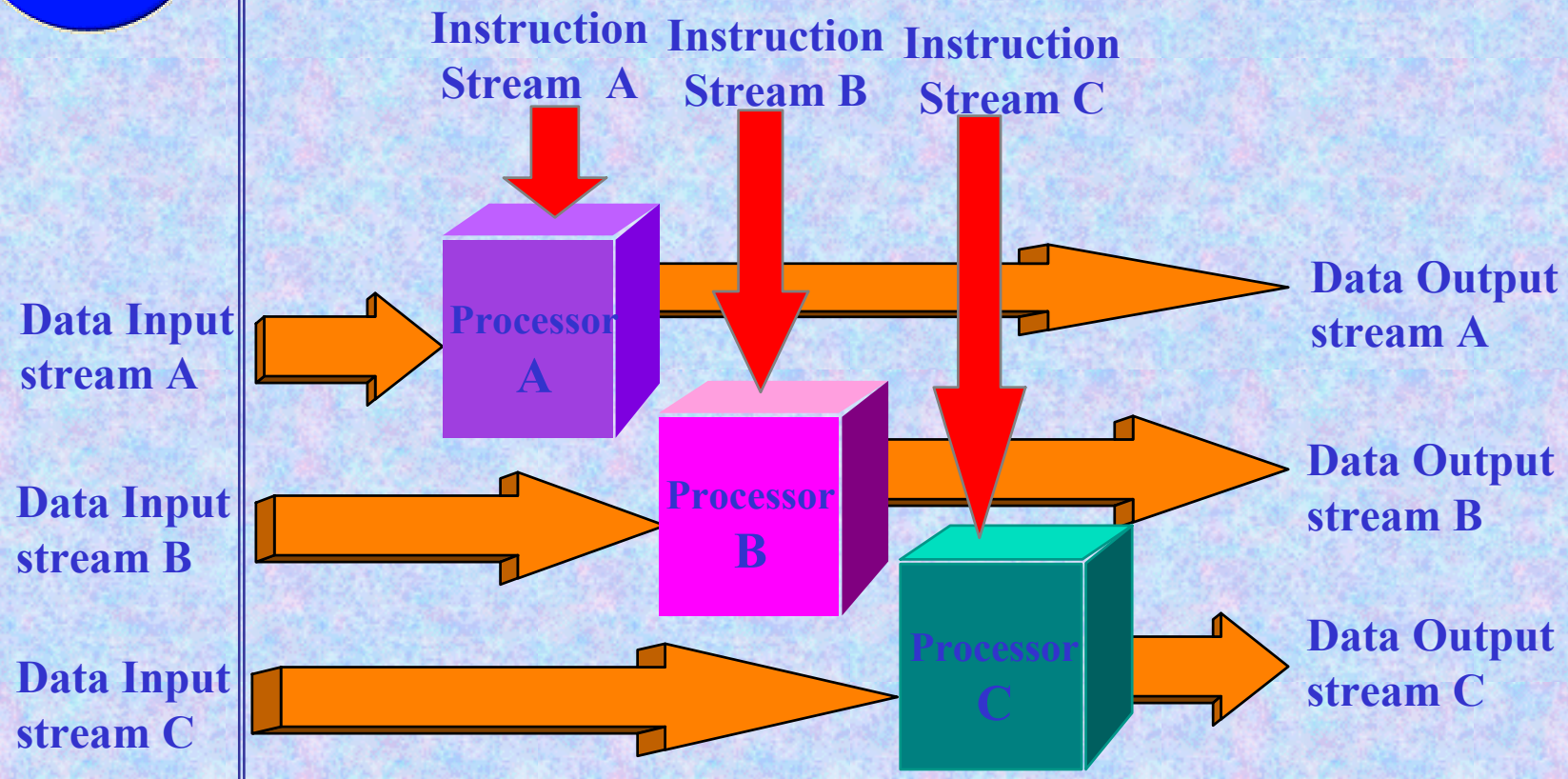
SIMD Architecture



$C_i <= A_i * B$

Ex: CRAY machine vector processing, Thinking machine cm*

MIMD Architecture

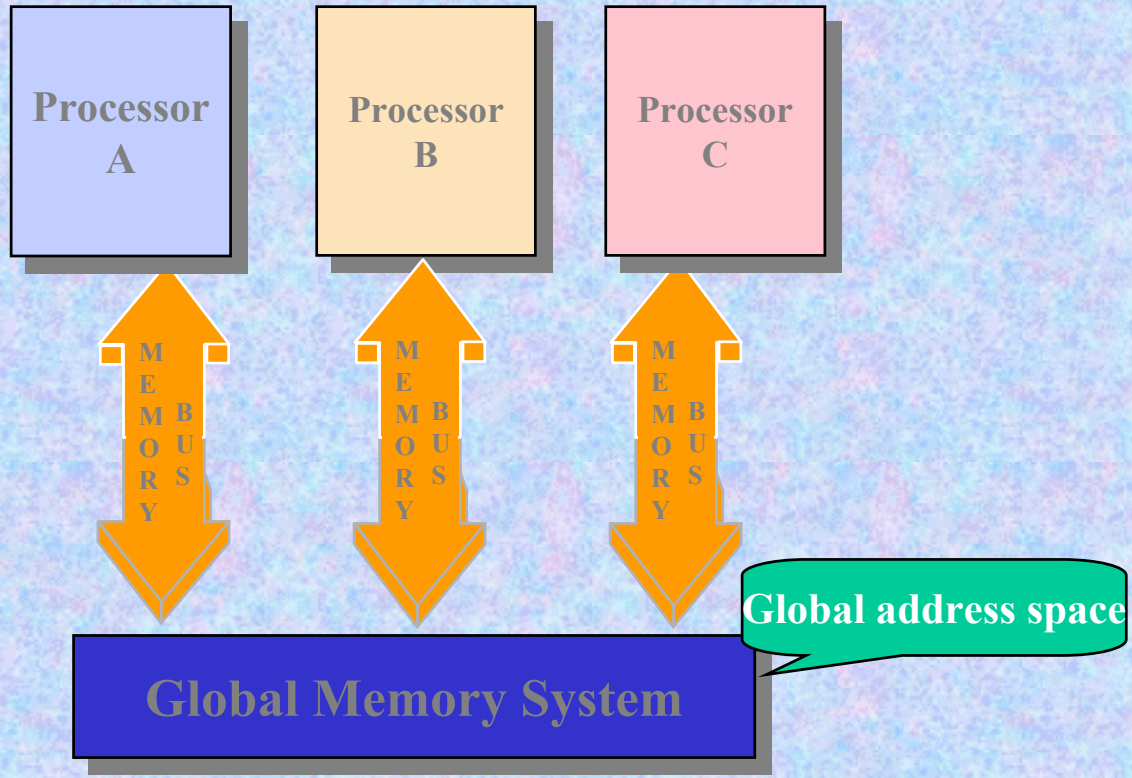


MIMD computer works asynchronously.

Shared memory (tightly coupled) MIMD

Distributed memory (loosely coupled) MIMD

Shared Memory MIMD machine



Comm: Source PE writes data to GM & destination retrieves it

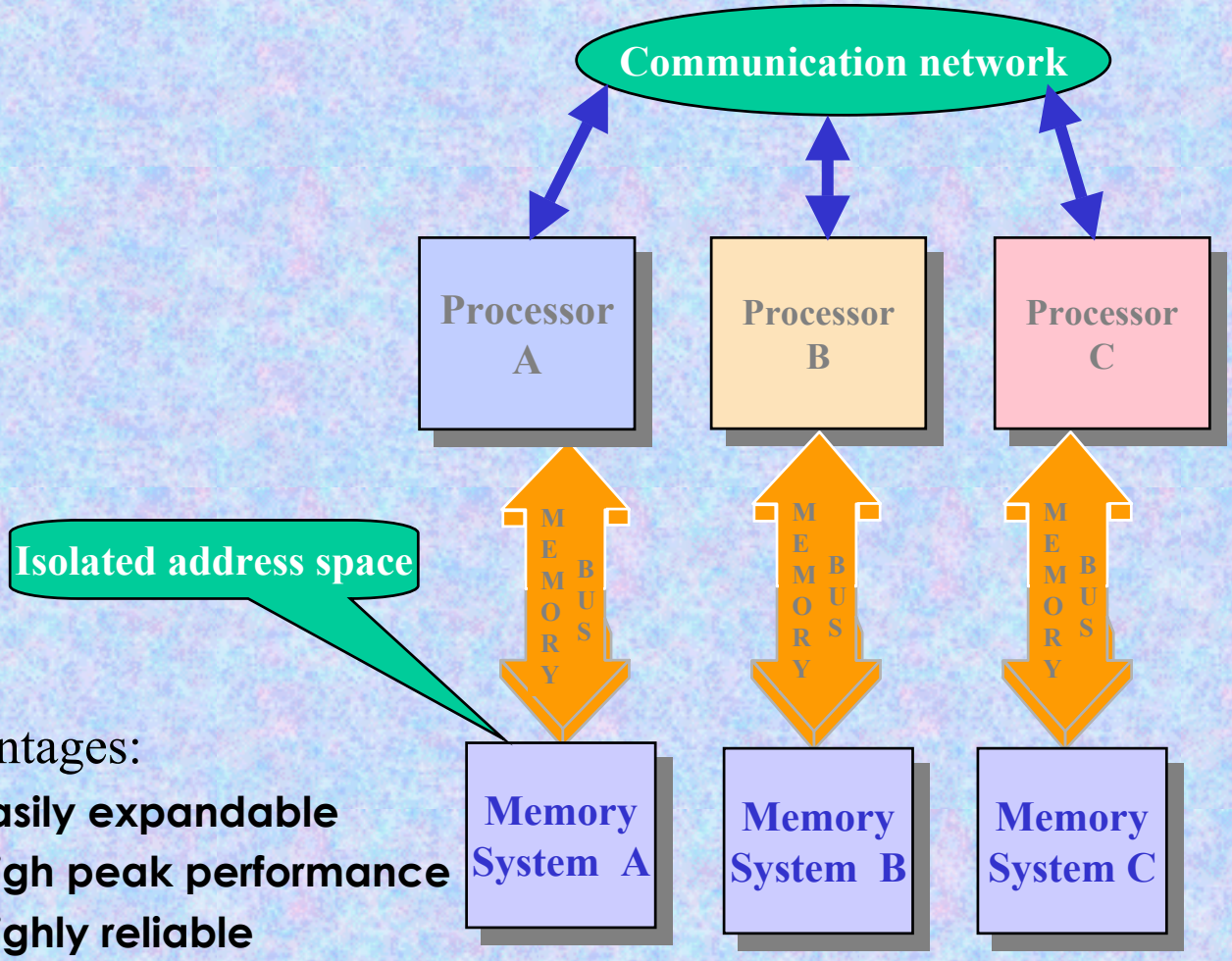
➔ Easy to program

➔ High sustained performance

➔ Limitation 1 : Increase of processors leads to memory contention.

➔ Limitation 2 : reliability & expandability. A memory component or any processor failure affects the whole system.

Distributed Memory MIMD



- Advantages:
  - ➔ Easily expandable
  - ➔ High peak performance
  - ➔ Highly reliable
- Drawbacks:
  - ➔ difficult load balancing, mapping
  - ➔ communication is more costly

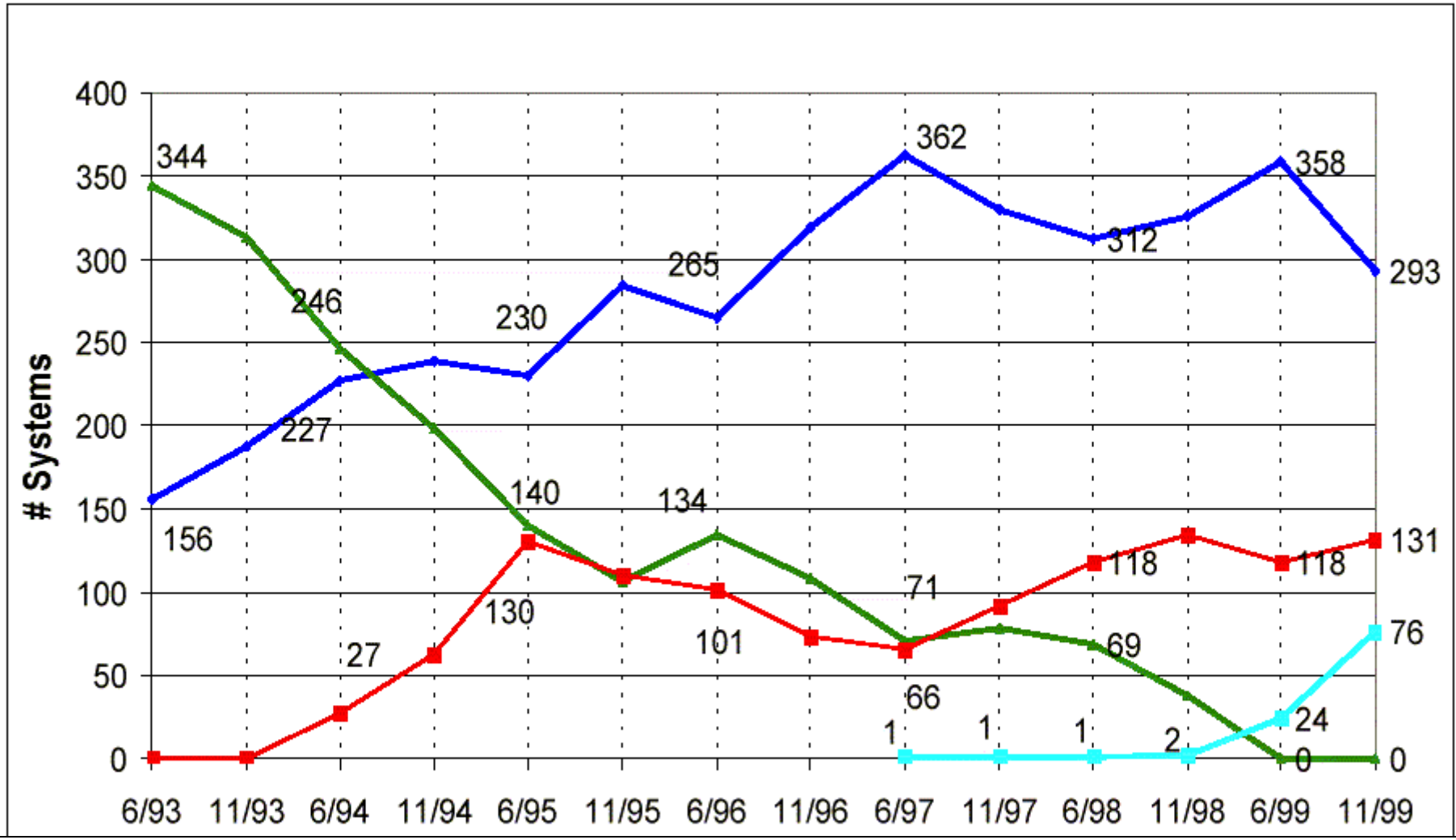
Main HPC Architectures

- SISD - mainframes, workstations, PCs.
- SIMD Shared Memory - Vector machines, Cray...
- MIMD Shared Memory - Sequent, KSR, Tera, SGI, SUN.
- SIMD Distributed Memory - DAP, TMC CM-2...
- MIMD Distributed Memory - Cray T3D, Intel, Transputers, TMC CM-5, plus recent workstation clusters (IBM SP2, DEC, Sun, HP).

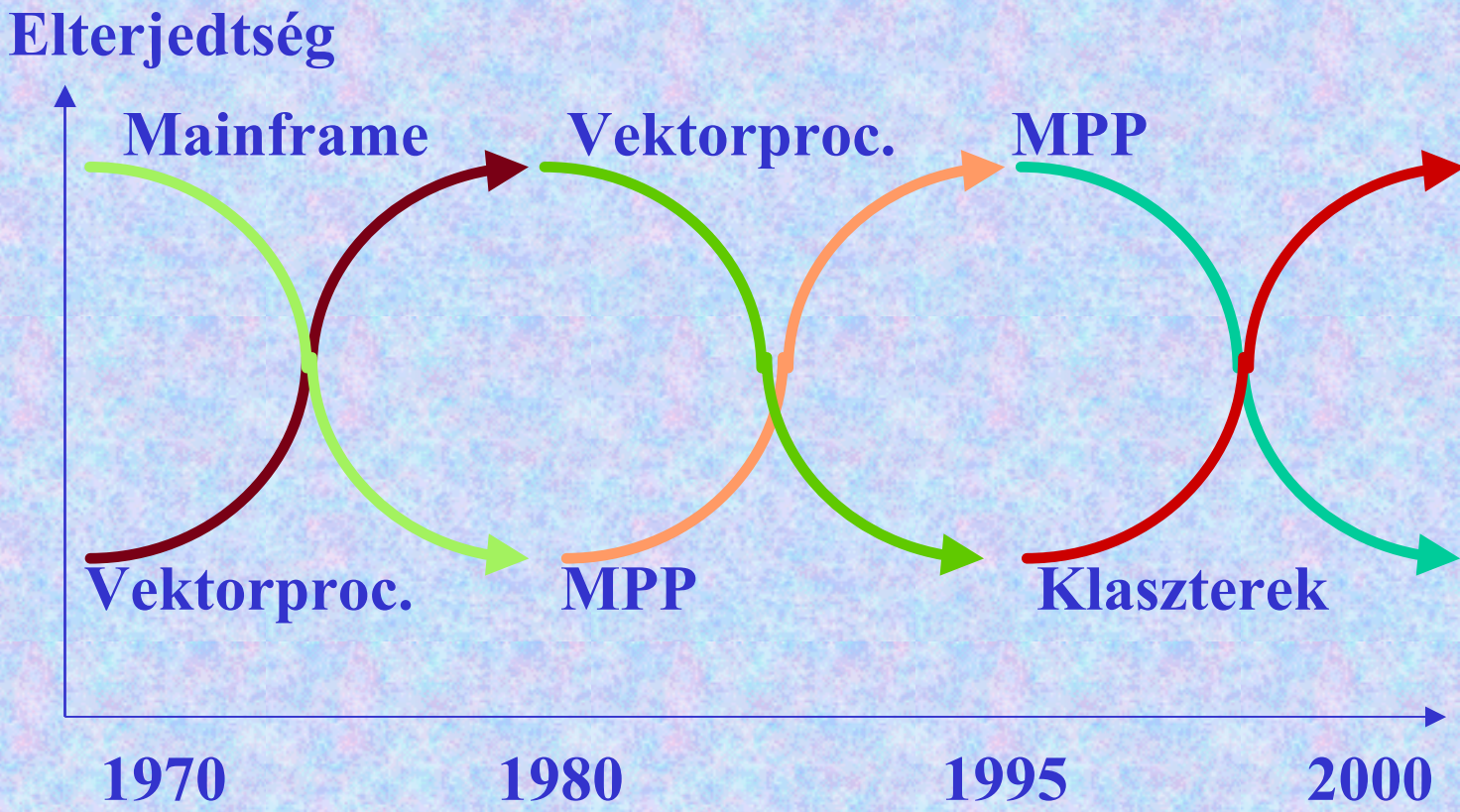
Main HPC Architectures

- NOTE: Modern sequential machines are not purely SISD - advanced RISC processors use many concepts from
  - vector and parallel architectures (pipelining, parallel execution of instructions, prefetching of data, etc) in order to achieve one or more arithmetic operations per clock cycle.

TOP 500 számítógép architektúrájának megoszlása

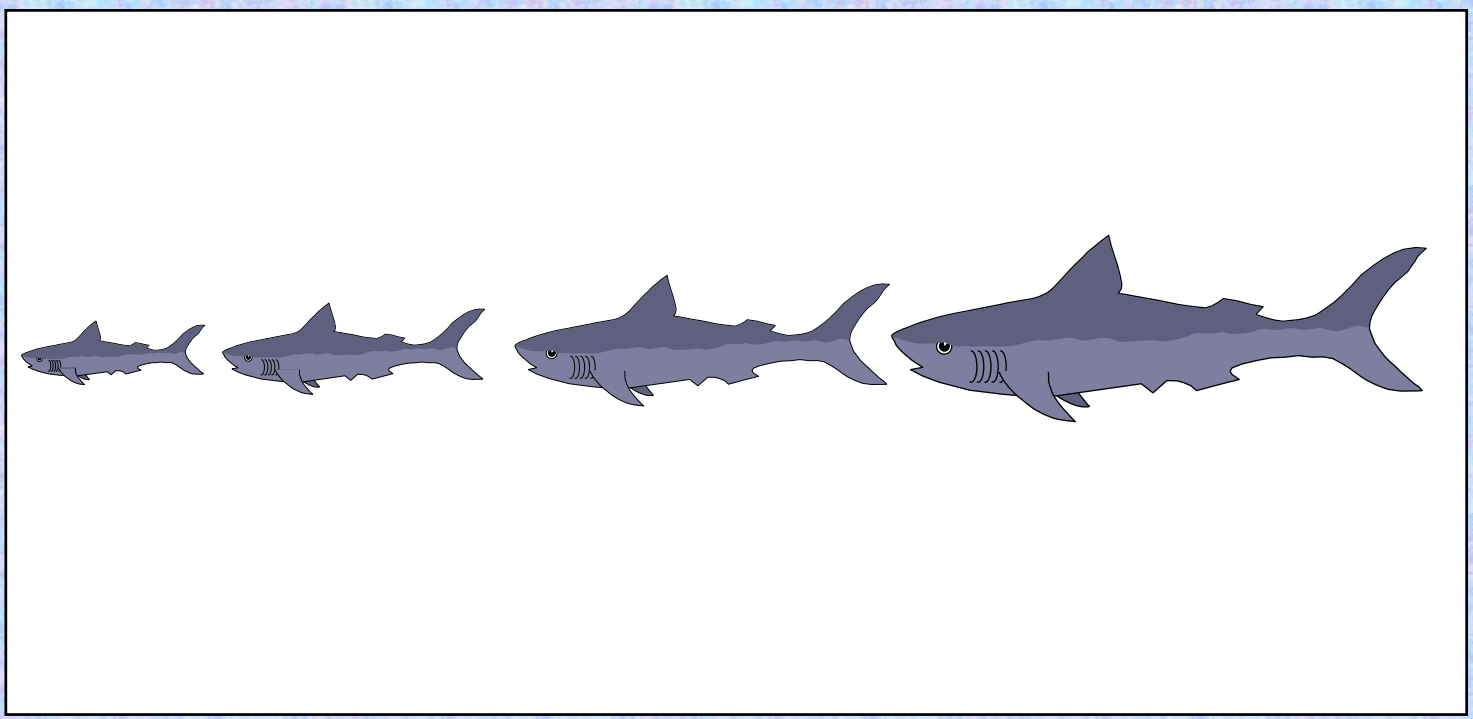


Szuperszámítógép típusok életciklusa

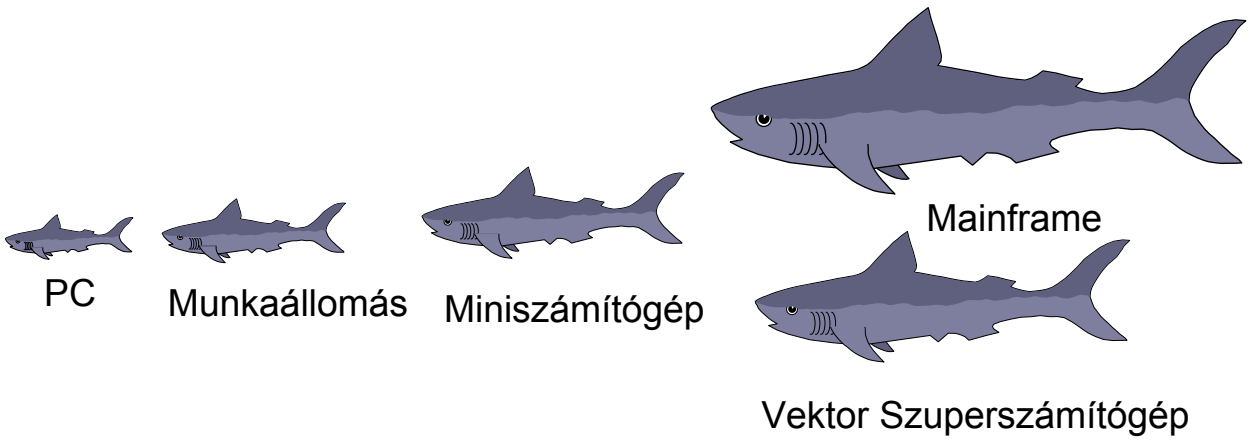


MPP: Masszívan Párhuzamos Processzorok

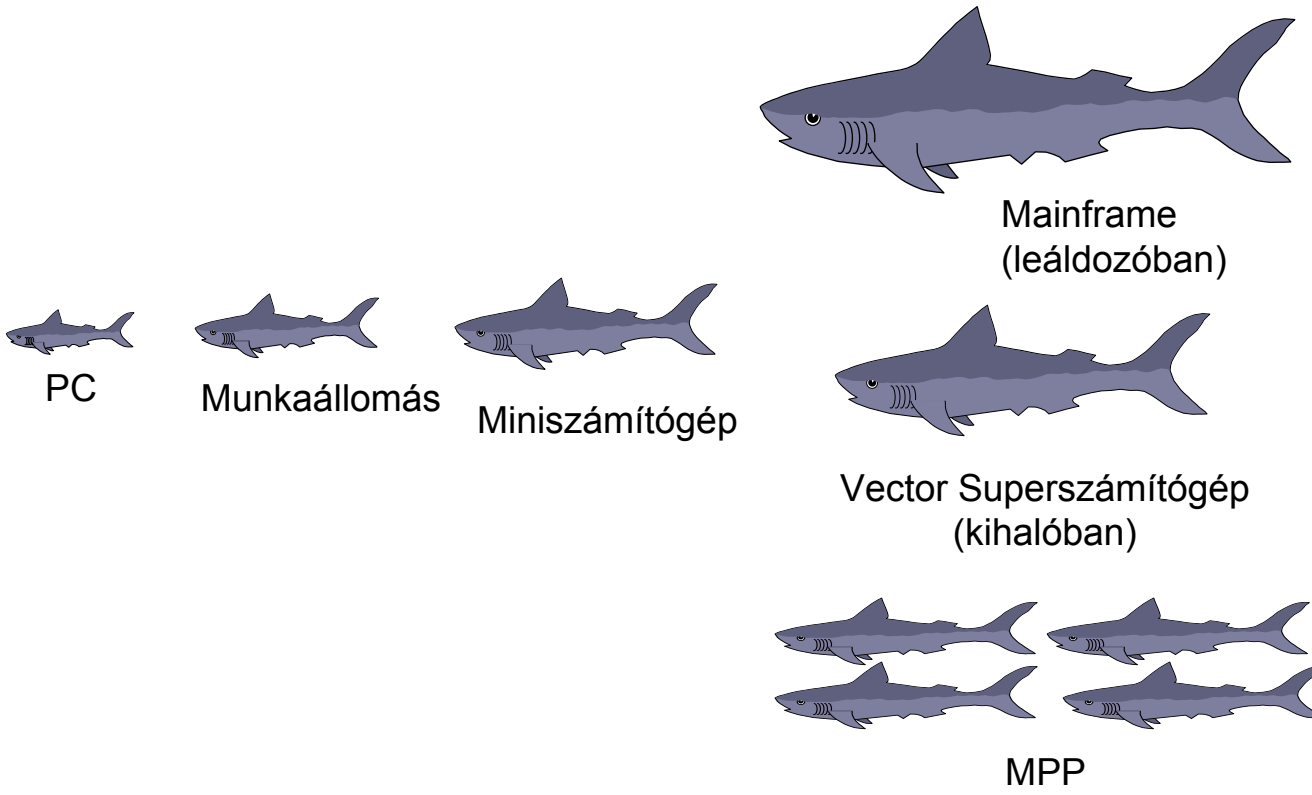
Halak táplálékklánc



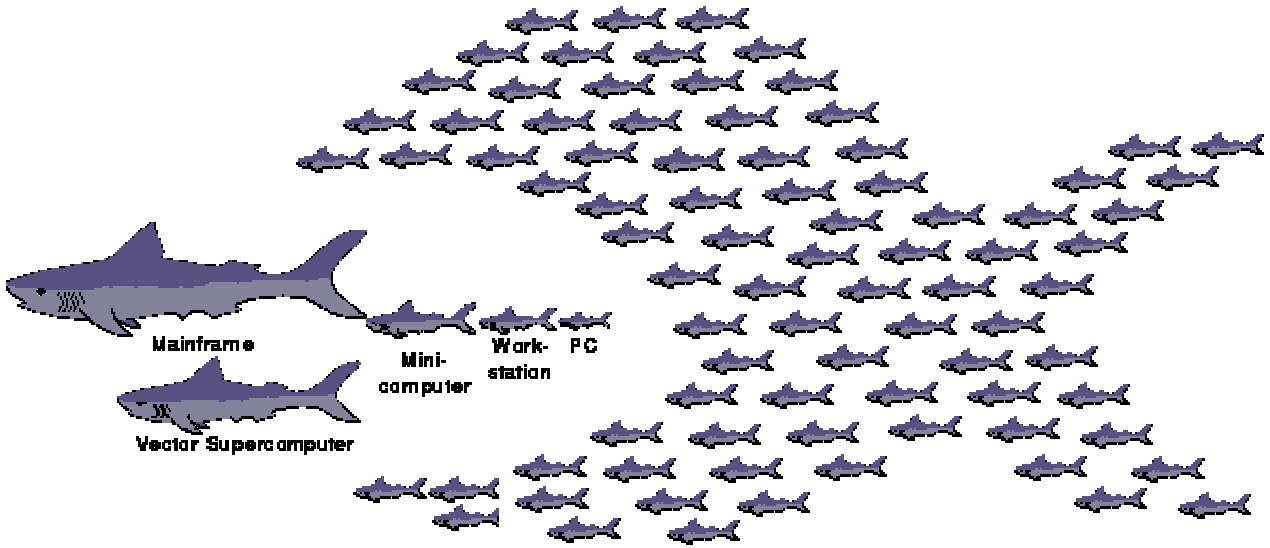
1984. évi számítógépek táplálékláncа



1994. évi számítógépek táplálékllánca

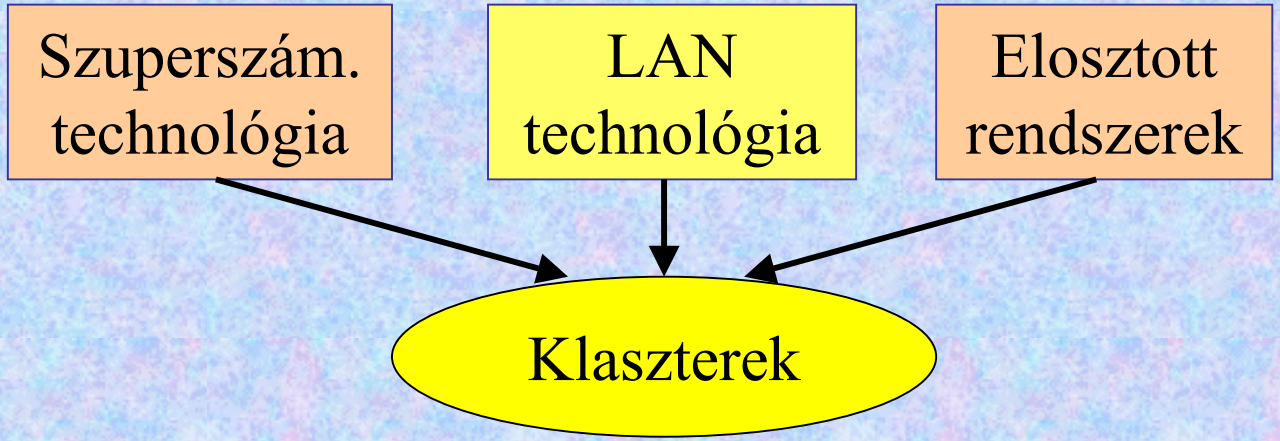


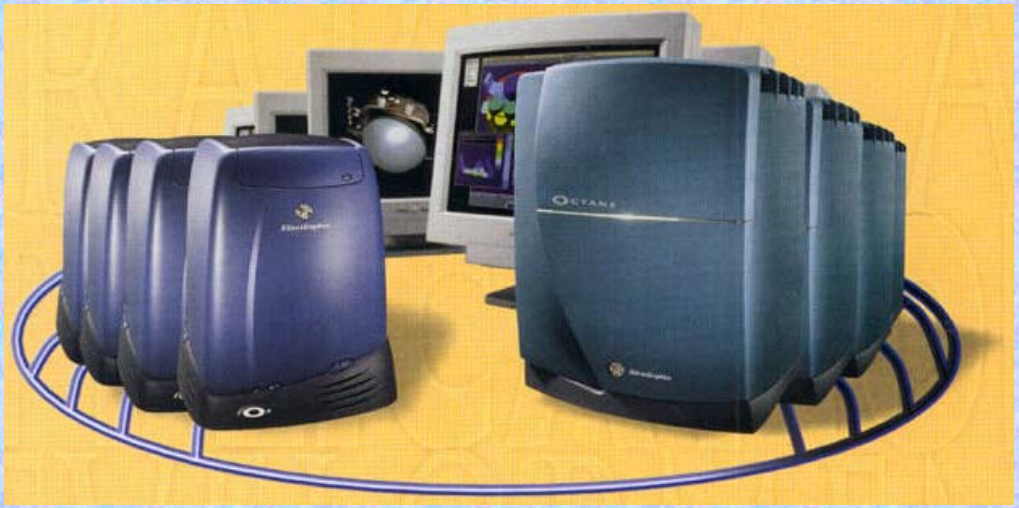
Jelen és jövő számítógépes táplálékláncа



Klaszter

A klaszter technológia eredete





A klaszterek három fő jellemzője

1. Egy klaszter **komplett számítógépeket** kapcsol össze (mint a metaszámítógépek)

2. A klasztert alkotó számítógépek **lazán csatoltak** (mint a metaszámítógépek)

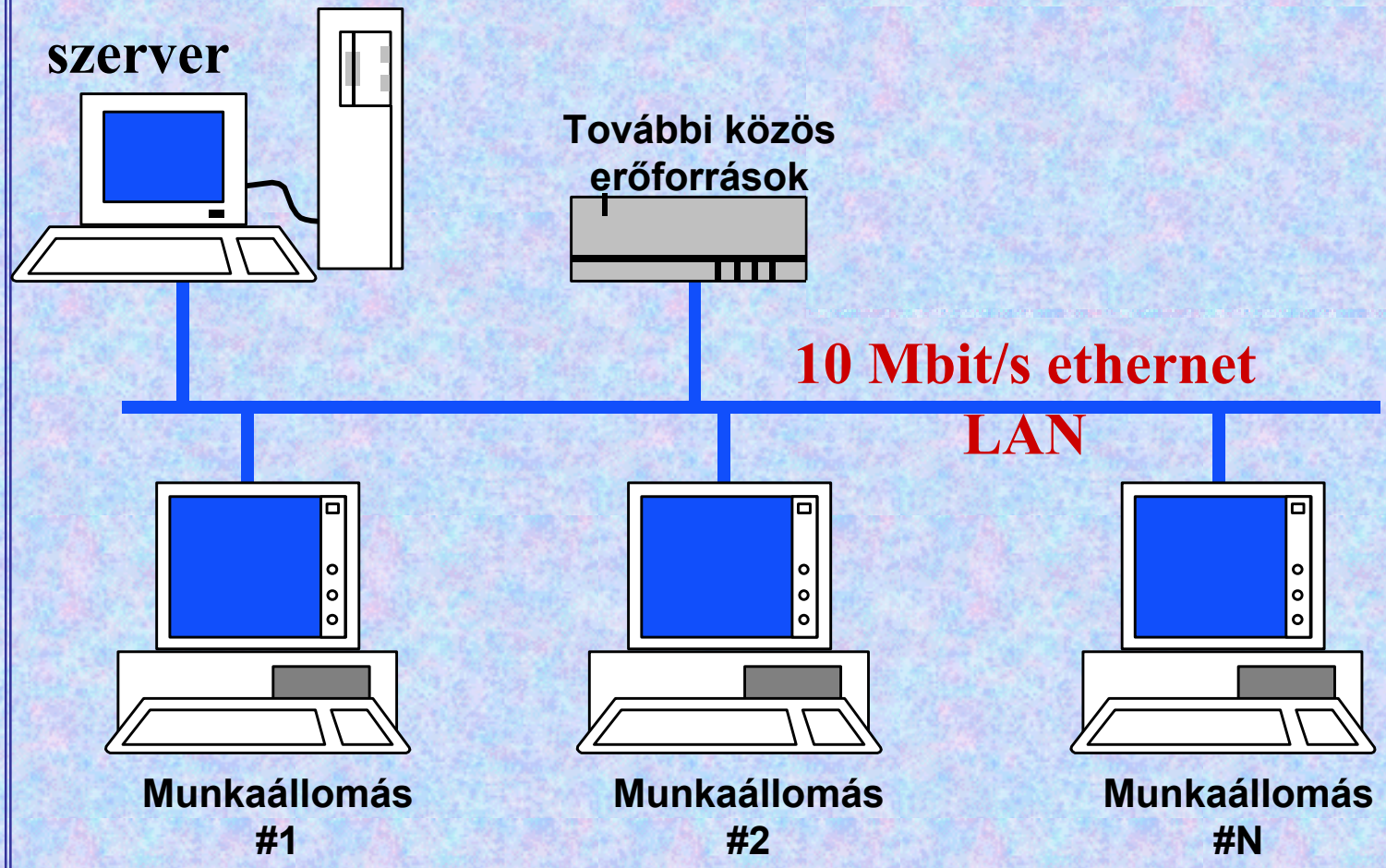
3. A klasztert mint egyetlen, egységes számítási erőforrást használjuk: **Single System Image** (mint a szuperszámítógépek)

A két alapvető klaszter típus

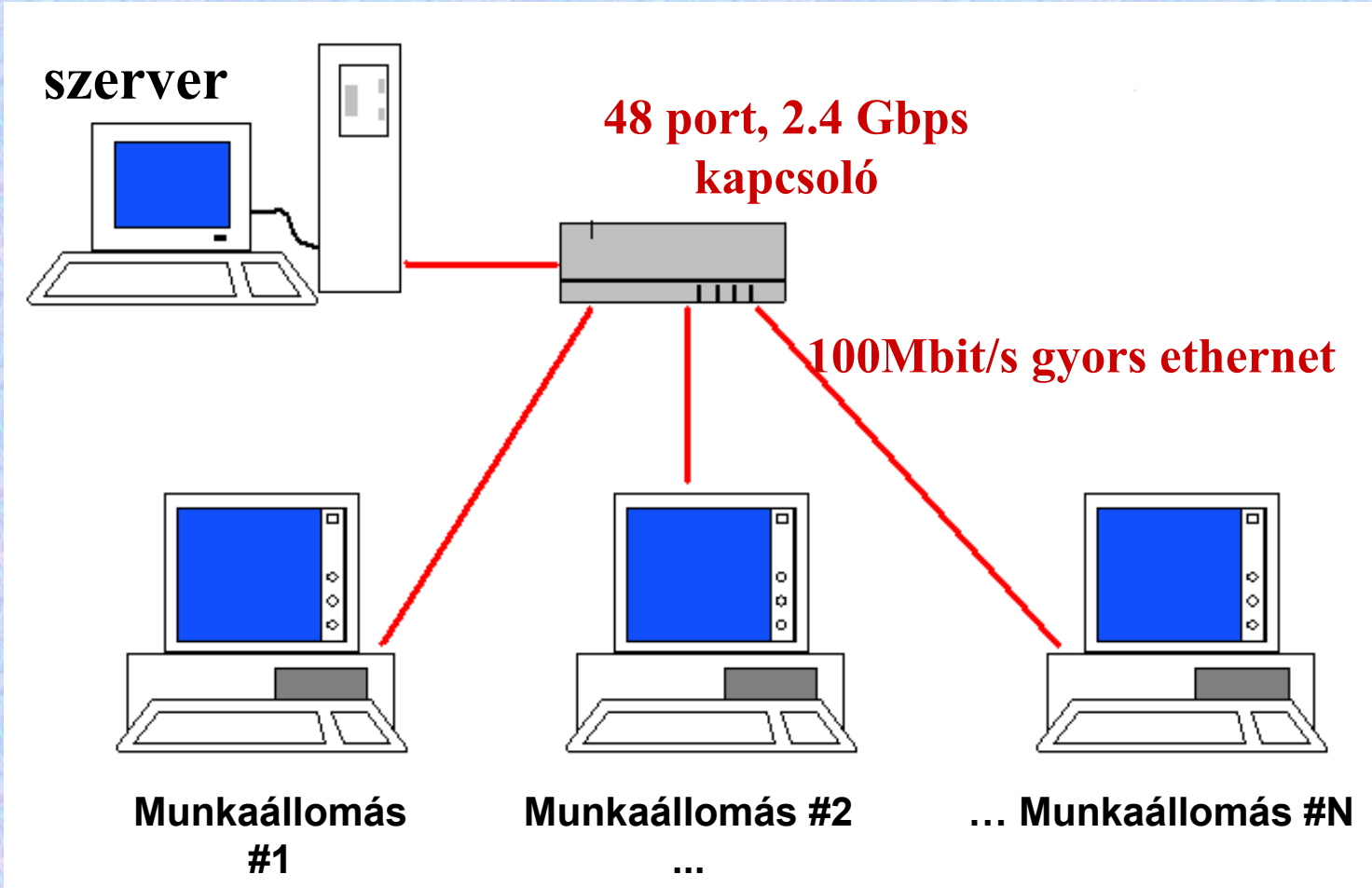
1. **Munkaállomás hálózat** (**NOW**: network of workstations)
   - **cél**: a szabad számítási ciklusok kihasználása
   - **módszer**: háttérben futó feladatok kiosztása
   - **tulajdonos**: munkaáll.-ok egyedi tulajdonban
   - **korlátozott SSI követelmények**

2. **Dedikált klaszterek** (szuperszám.-ek)
   - **cél**: nagy teljesítmény és sebesség
   - **módszer**: párhuzamos feldolgozás
   - **tulajdonos**: közös tulajdon
   - **erős SSI követelmények**

Tipikus NOW struktúrája



Tipikus dedikált klaszter struktúrája



Műszaki jellemzők

- 1 db DELL Precision 610M szerver gép
  - 2db Intel Pentium III Xeon 550MHz processzor
  - 256 MB hibajavítós SDRAM memória
  - 2x 18GB Ultra2 SCSI diszk
  - 3db 100Mbit-es Ethernet interfész kártya
  - 3D gyorsító videókártya, 32 MB RAM
  - 40-szeres sebességű SCSI CD-ROM olvasó
  - 17” DELL monitor

Műszaki jellemzők

- 28 db DELL Precision 410M munkaállomás
  - 2db Intel Pentium III 500MHz processzor
  - 128 MB hibajavítós SDRAM
  - 9.1 GB Ultra2 SCSI diszk
  - 100Mbit-es Ethernet interfész kártya
  - 3D gyorsító videókártya, 32 MB RAM
  - 40-szeres sebességű SCSI CD-ROM olvasó
  - 15” DELL monitor
- Hálózat: 100Mbit-es Ethernet
  - 48 portos Cisco 100Mbit-es Ethernet kapcsoló (full-duplex, 2.4Gbps)

Az MTA SZTAKI klaszter jellemzői

- 29 db duál-Pentiumos PC
- Szuperszámítógép kapacitás
  - Teljes memória kapacitás: 3.84 GB
  - Teljes diszk kapacitás: 290 GB
  - Hálózati áteresztőképesség: 2.4 Gbps
- Magyarország egyik leggyorsabb számítógép konfigurációja
  - Csúcssebesség: ~ 30 Gflop
  - Top500 belépő: 60 Gflop

Példa klaszter: Berkeley NOW

- 100 Sun UltraSparcs
  - 200 disks
- Myrinet SA
  - 160 MB/s
- Fast comm.
  - AM, MPI, .
- Ether/ATM switched external net
- Global OS
- Self Config

Szuperszámítógépek, klaszterek és metaszámítógépek összehasonlítása I.

	Supercomputer	Cluster	NOW	Metacomputing system
Processing units (nodes)	Microprocessors	PCs, workstations	PCs, workstations	Supercomputers, clusters, PCs, workstations
Number of nodes	100 - 1000	10 - 100	10 - 100	100 - 10000
Communication network	Buses, switches	LAN	LAN	Internet
Node OS	Homogeneous	Typically homogeneous	Typically heterogeneous	Heterogeneous
Inter-node security	Nonexistent	Rarely required	Necessary	Necessary

Szuperszámítógépek, klaszterek és metaszámítógépek összehasonlítása II.

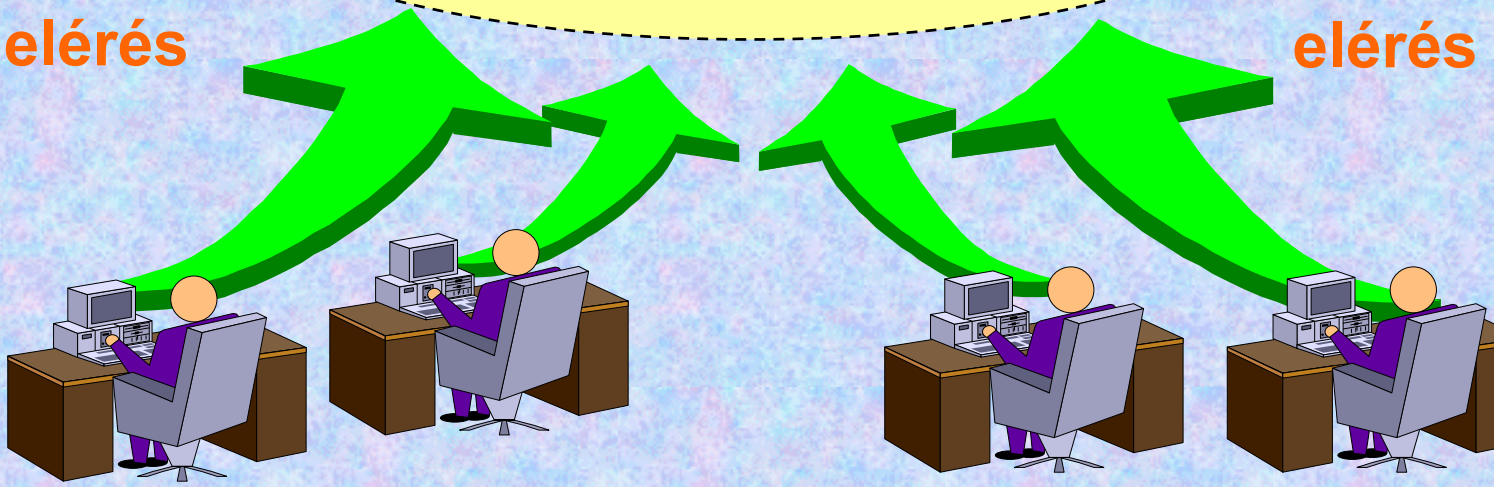
	Supercomputer	Cluster	Metacomputing system
Programming models	• Shared memory • Message passing	• Shared memory • Message passing • Peer-to-peer • Client-server	• Message passing • Client-server • Code shipping • Proxy computing • Intelligent mobile agents
Programming language	• HPF • (C/Fortran)+MPI	• HPF • (C/Fortran)+MPI	• HPF • (C/Fortran)+MPI • Java/CORBA
Middleware	• No	• Limited forms	• Toolkit approach • Three-tier commodity (Java/CORBA) • Object-oriented
Programming environment	• Toolkit approach • Integrated environment	• Toolkit approach • Integrated environment	• Toolkit based • Application specific • Integrated environment
Resource allocation	• Mapping • Load balancing	• Mapping • Load balancing	• Resource manager
QoS	No	No	Yes

Klaszter elérési üzemmódok

Klaszter = Közös erőforrás (Processzorok, Memória, Diszkek)

Lokális elérés

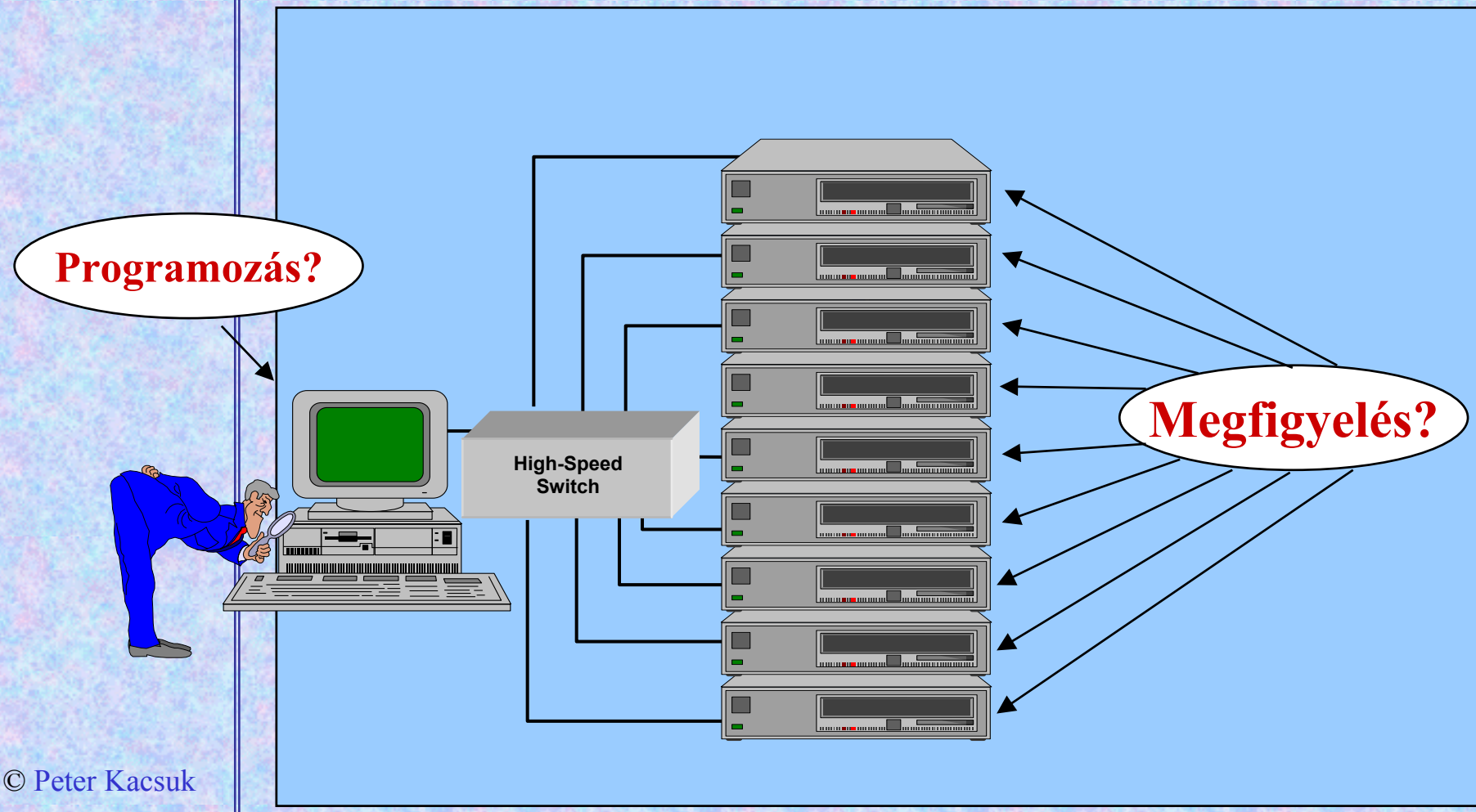
Interne elérés



Sok felhasználó számára garantál legalább egy munkaállomást

Sok munkaállomást biztosít néhány felhasználó számára

Klaszterek programozási nehézségei



Szoftver környezet a klaszteren

- Operációs rendszerek
  - RedHat Linux 6.1
  - Microsoft NT
- Kommunikációs könyvtárak párhuzamos programozáshoz
  - PVM (Parallel Virtual Machine)
  - MPI (Message Passing Interface)
- Virtuális közös memóriás rendszerek
  - OpenMP

- Párhuzamos programfejlesztő rendszer
  - **P-GRADE**

Köszönöm a figyelmüket

További információ: www.lpds.sztaki.hu