Küme Bilgisayarlar

Onur Temizsoylu

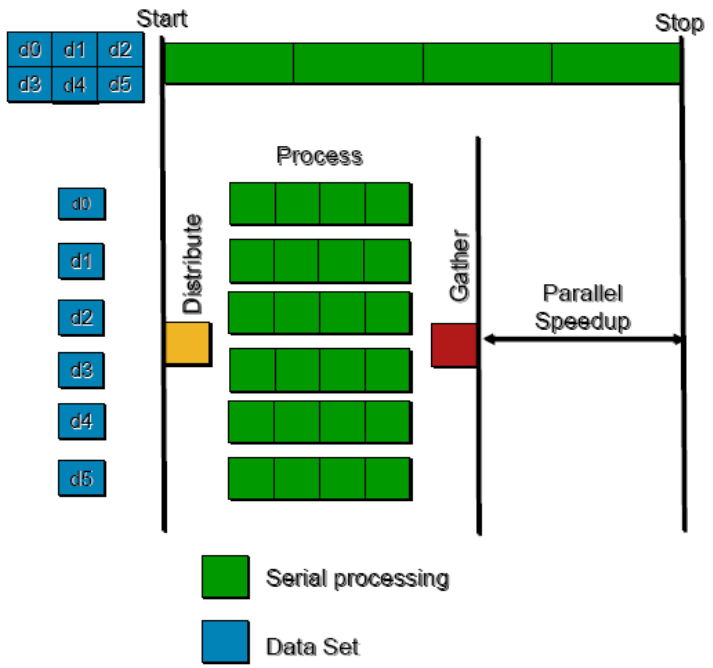
Bilgisayar Mühendisliği, ODTÜ, Ankara
04/02/2008

İçerik

- Terminoloji
- Paralelleştirme, Paralel Donanım Mimarileri
- Neden hesaplamada kümeleme?
- Kümeleme nedir?
  - Yüksek kullanılabilirlik kümeleri
  - Yük dengeleme kümeleri
  - Veritabanı kümeleri
  - Yüksek başarımlı hesaplama kümeleri
- YBH küme mimarileri
  - Sunucular
  - Ağ bağlantısı
  - Depolama
  - İşletim sistemi
  - Orta katman yazılımları
  - Uygulama programlama arayüzleri

Terminoloji

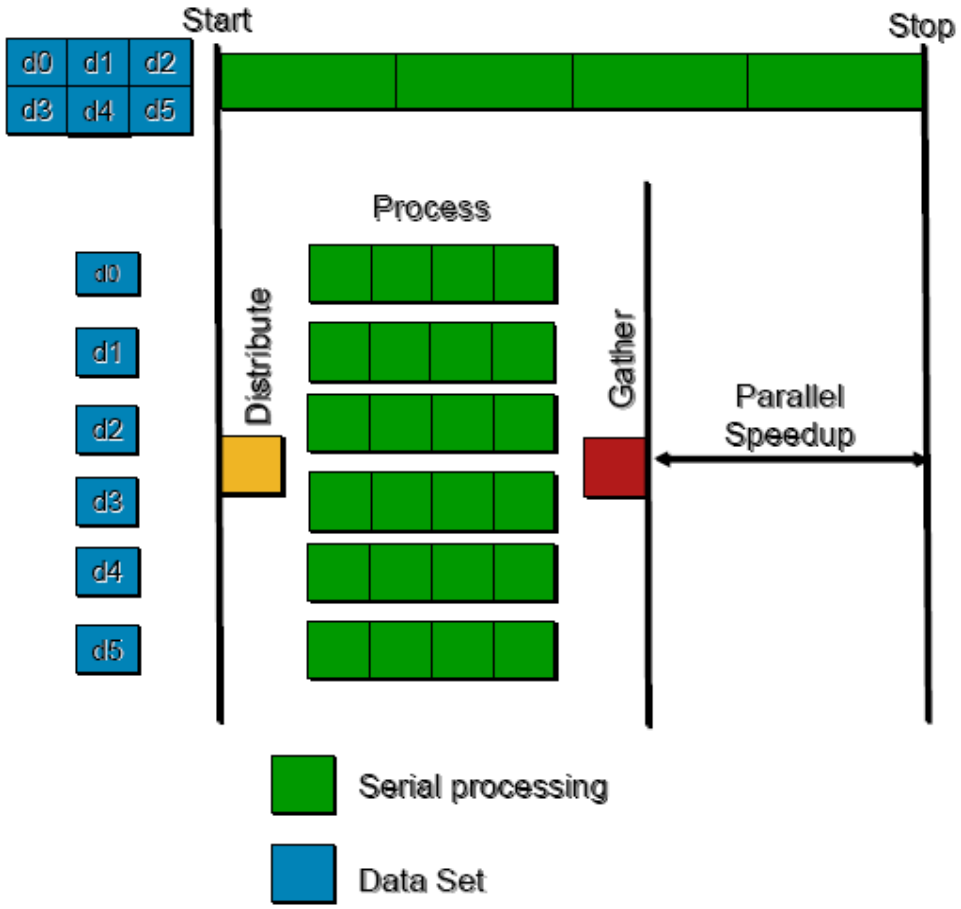
- Süreç (“Process”)
- İş Parçacığı (“Thread”)
- Görev (“Task”)



- Hızlanma (“Speedup”)
- Ölçeklenebilirlik (“Scalability”)
- Verimlilik
- Senkronizasyon (“Synchronization”)
- Paralel Ek Yükü (“Parallel Overhead”)
- Süperbilgisayar

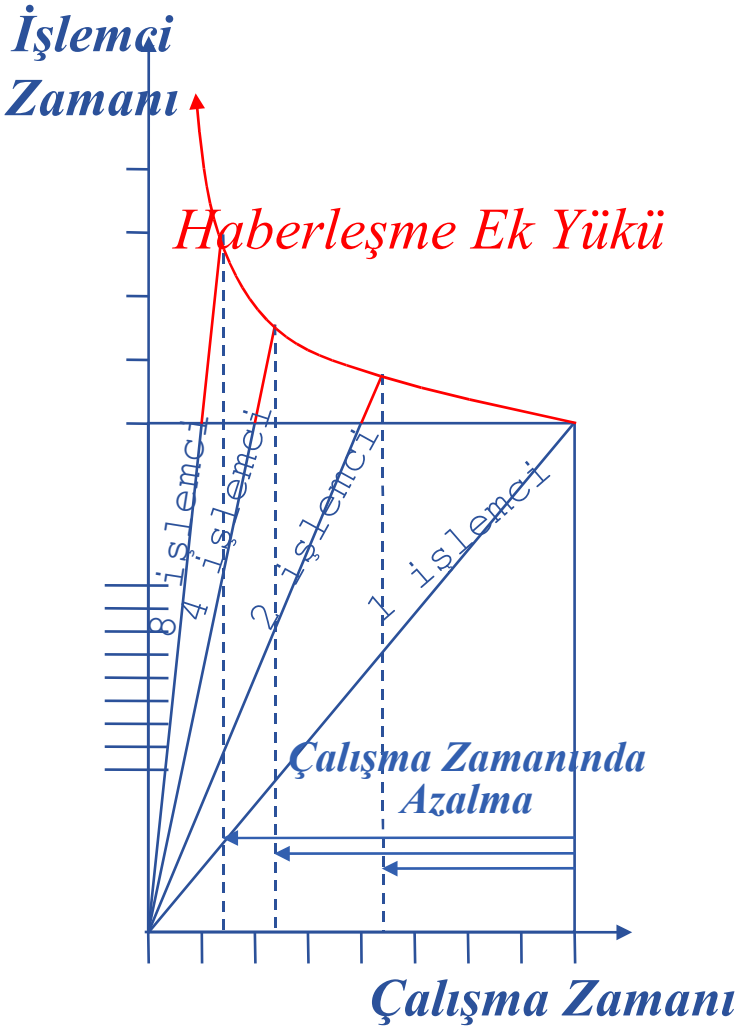
Paralelleştirme

– Bir işin paralelleştirilmesinde programın toplam çalışma zamanını azaltmak amaçlanır.



Paralelleştirme Ek Yükü

- Ek yük:
  - İşlemcilerde fazladan geçen süre
  - İletişim ek yükü
  - Senkronizasyon ek yükü
  - Programın paralel olmayan/ olamayan parçaları
- Paralel programlamada ek yük ve çalışma zamanı hızlanma ve verimlilik ile ifade edilir.



Hızlanma ve Verimlilik

- İ sayıda işlemcide programın toplam işlemci zamanını Z(i) olarak ifade edelim.

Hızlanma (i) = Z(1) / Z(i)

Verimlilik (i) = Hızlanma (i) / i

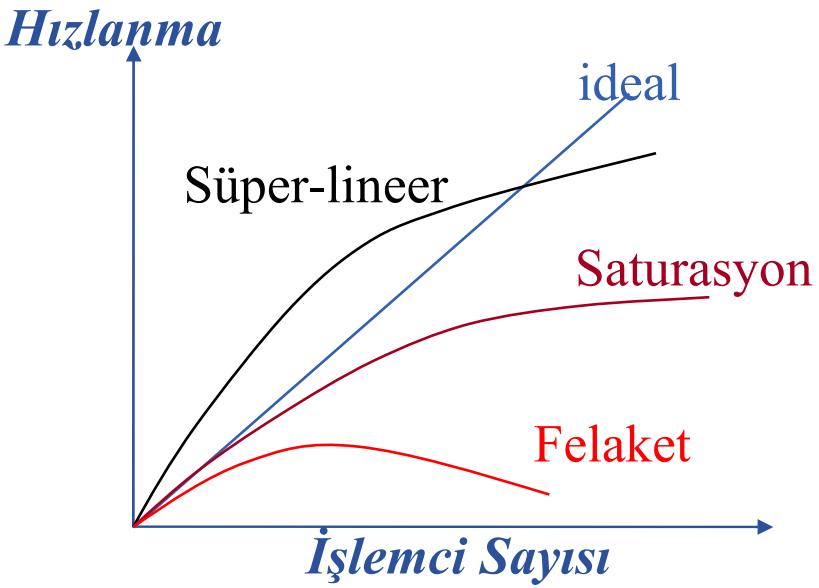
- İdeal durumda:

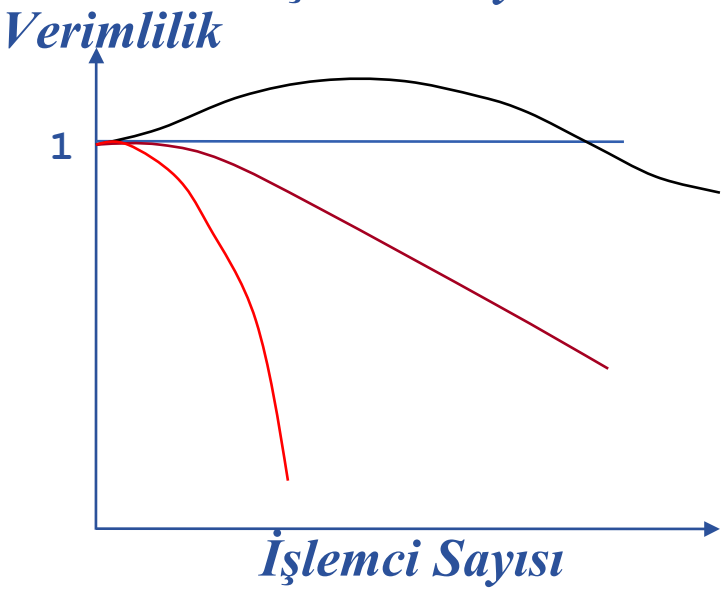
Z (i) = Z (1) / i

Hızlanma (i) = i

Verimlilik (i) = 1

- Ölçeklenebilir programlar büyük işlemci sayılarında bile verimli kalırlar.





Amdahl Yasası

- Amdahl yasası:
  - “Kodun paralel olmayan kısmı (ek yük), kodun ölçeklenebilirliği konusunda üst limiti oluşturur.”
- Kodun seri kısmını s, paralel kısmını p olarak ifade edersek:

$1 = s + p$

$Z(1) = Z(s) + Z(p)$

$= Z(1) * (s + p)$

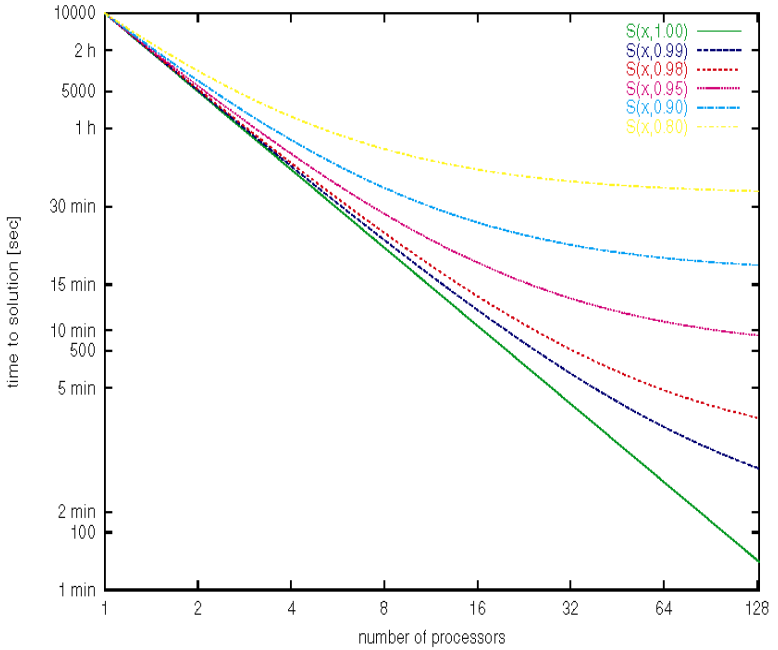
$= Z(1) * (p + (1-p))$

$Z(i) = Z(1) * (p/i + (1-p))$

Hızlanma (i) $= Z(1) / Z(i)$

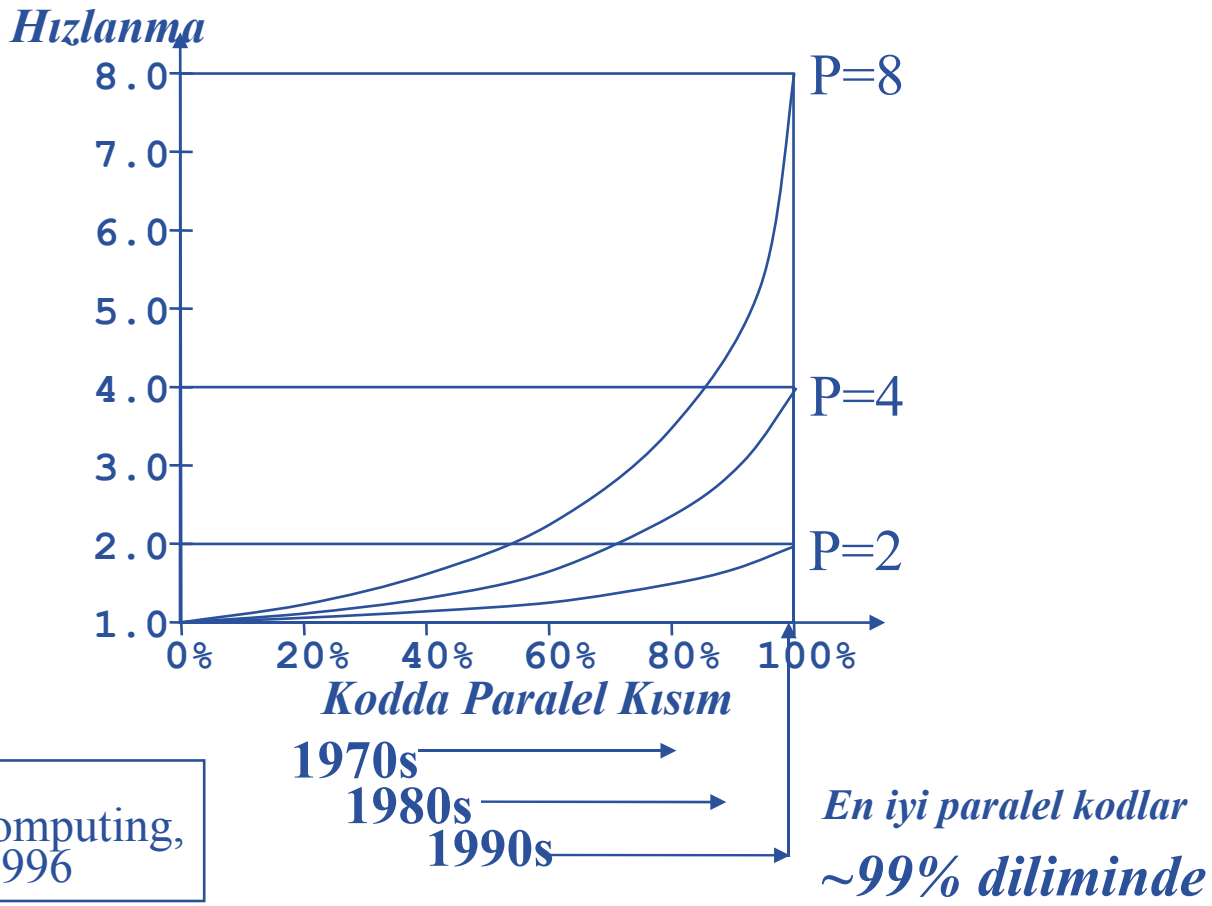
$= 1 / (p/i + 1 - p)$

Hızlanma (i) $< 1 / (1 - p)$



Pratikte

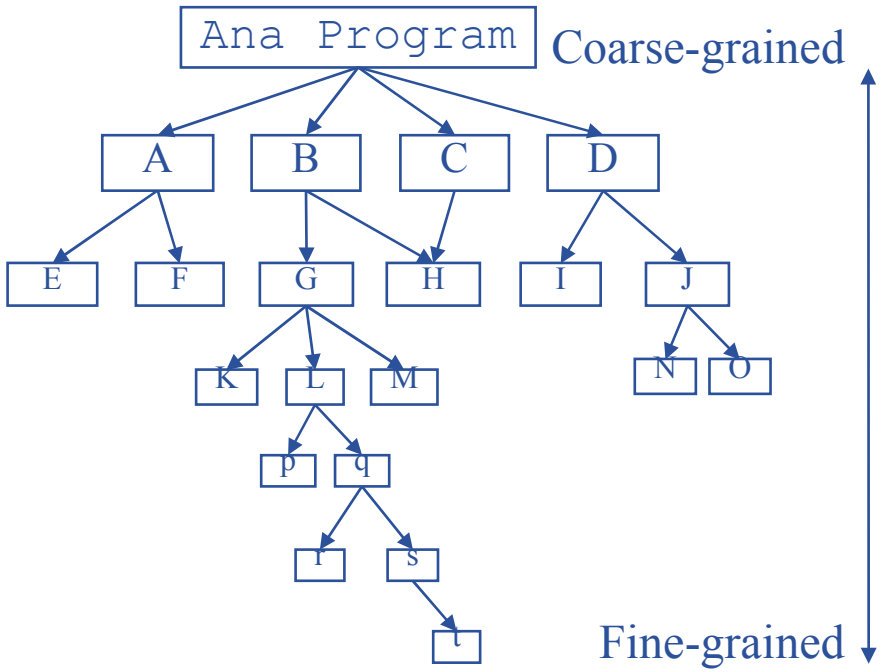
- Pratikte programları paralelleştirmek Amdahl yasasında görüldüğü kadar zor değildir.
- Ancak programın çok büyük bir kısmını paralel işlem için harcaması gereklidir.



David J. Kuck,
Hugh Performance Computing,
Oxford Univ.. Press 1996

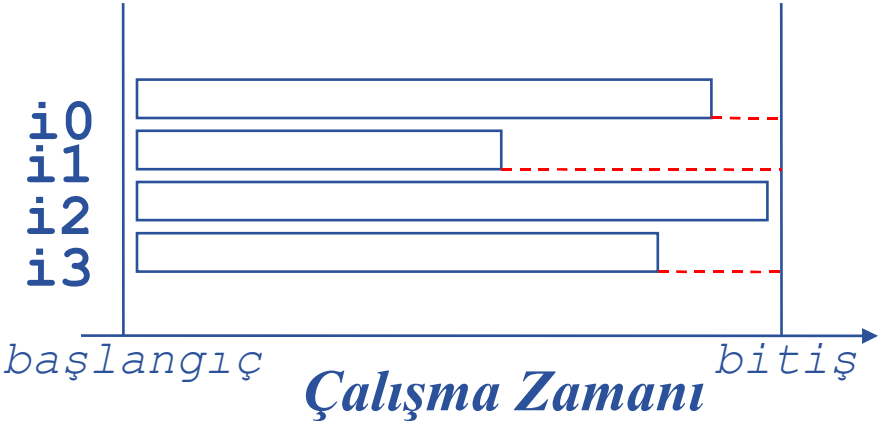
Coarse/Fine Grained Paralel

- Fine-Grained:
  - Genelde her döngüde paralelleştirme vardır.
  - Çok sayıda döngü paralleleştirilir.
  - Kodun çok iyi bilinmesine gerek yoktur.
  - Çok fazla senkronizasyon noktası vardır.
- Coarse-Grained:
  - Geniş döngülerle paralleştirme yapılır.
  - Daha az senkronizasyon noktası vardır.
  - Kodun iyi anlaşılması gerekir.

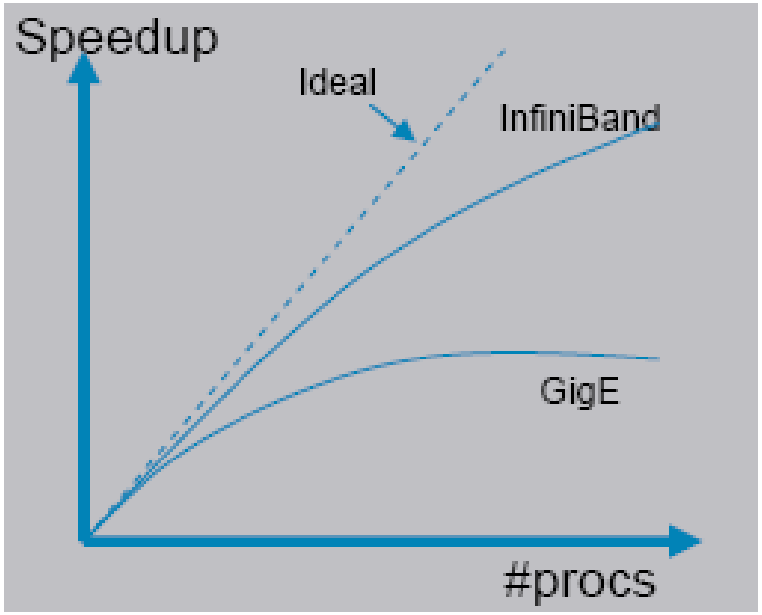


Ölçeklenebilirlik

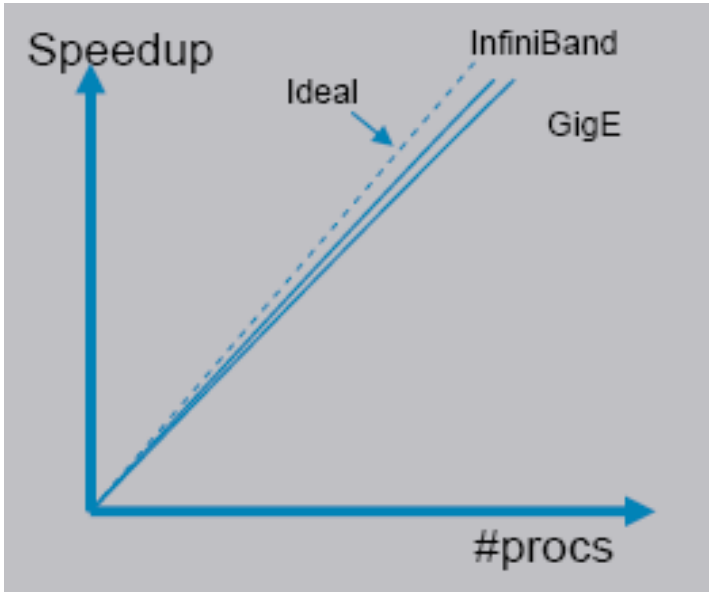
- Ölçeklenebilirliği etkileyen diğer faktörler:
  - **İş parçacıkları arası yük dengesizliği** : Bir kodun herhangi bir paralel kısmının çalışma zamanı en uzun süren iş parçacığının çalışma zamanıdır. Coarse-Grained programlamada ortaya çıkması daha olasıdır.
  - **Çok fazla senkronizasyon:** Kodda küçük döngüler sırasında her seferinde senkronizasyon yapılırsa bu ek yük getirir. Fine-Grained programlamada ortaya çıkması daha olasıdır.



- Sıkı bağlı sistemler:
  - Süreçler arasında yoğun haberleşme
  - Gecikme süresine hassas
  - Ortak Bellek Paralel
  - Dağıtık Bellek Paralel



- Gevşek bağlı sistemler:
  - Süreçler arasında haberleşme azdır veya hiç yoktur.
  - Gecikme süresine hassas değillerdir. Ancak bant genişliği veri transferi için etkili olabilir.
- Parametrik çalışan uygulamalar
  - Süreçler arasında haberleşme yoktur.
  - Kümelerde, grid altyapılarında çalışan uygulamaların çoğunluğunu oluştururlar.

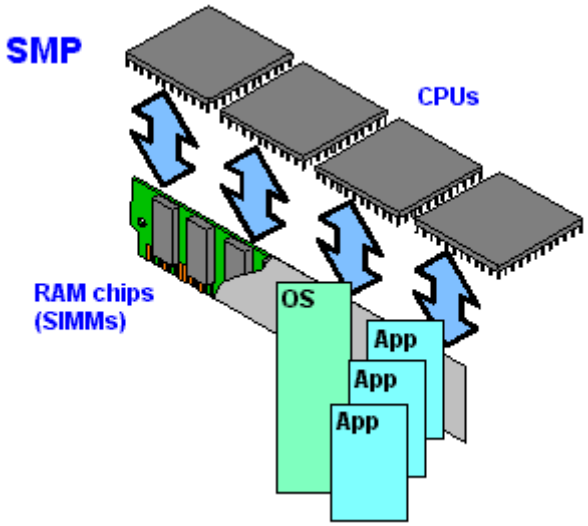


Paralel Donanım Mimarileri

- SMP makineler
- MPP makineler
- NUMA makineler
- Superscalar işlemciler
- Vektör makineler
- Küme bilgisayarlar

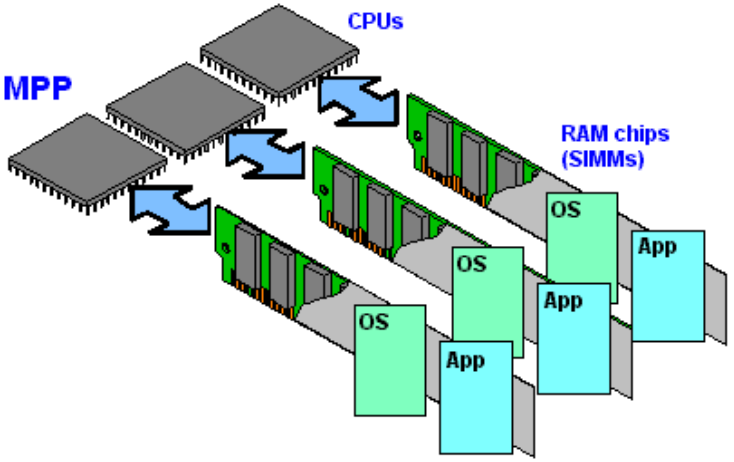
SMP

- SMP, birden fazla eş işlemcinin ortak bir belleğe bağlandığı çok işlemcili bir bilgisayar mimarisidir.
- SMP sistemler, görevleri işlemciler arasında paylaşabilirler.
- SMP sistemler, paralel hesaplama için kullanılan en eski sistemlerdir ve hesaplamalı bilimlerde yoğun bir şekilde kullanılırlar.



MPP

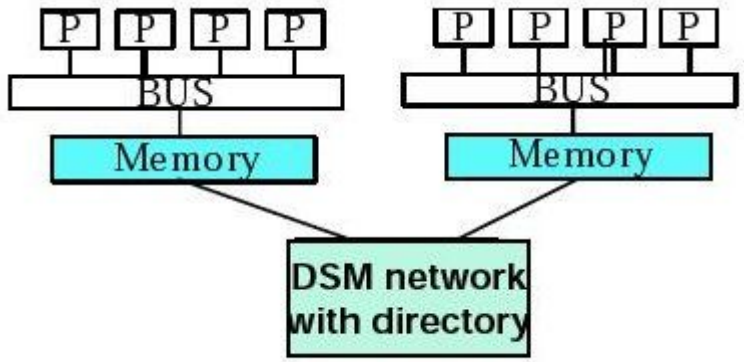
- MPP, binlerce işlemci kullanılabilen çok işlemcili bir mimaridir.
- Bir MPP sisteminde her işlemci kendi belleğine ve işletim sistemi kopyasına sahiptir.
- MPP sistemler üzerinde çalışacak uygulamalar eş zamanda çalışacak eş parçalara bölünebilmelidirler.
- MPP sistemlere yeni işlemci ekledikten sonra uygulamalar yeni paralel kısımlara bölünmelidirler. SMP sistemler ise bundan çok iş parçacığı çalıştırabilir yapıları sayesinde hemen faydalanırlar.

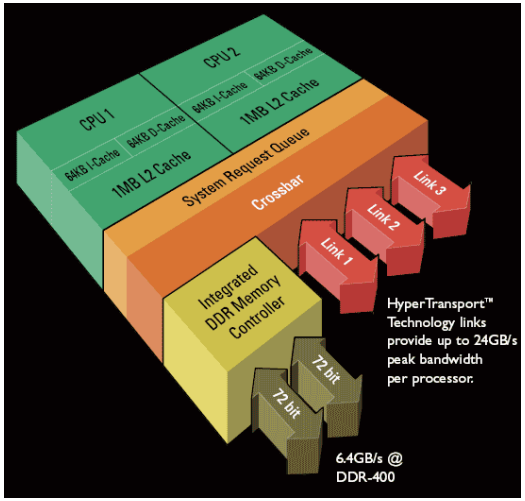


From Computer Desktop Encyclopedia
© 1998 The Computer Language Co. Inc.

NUMA

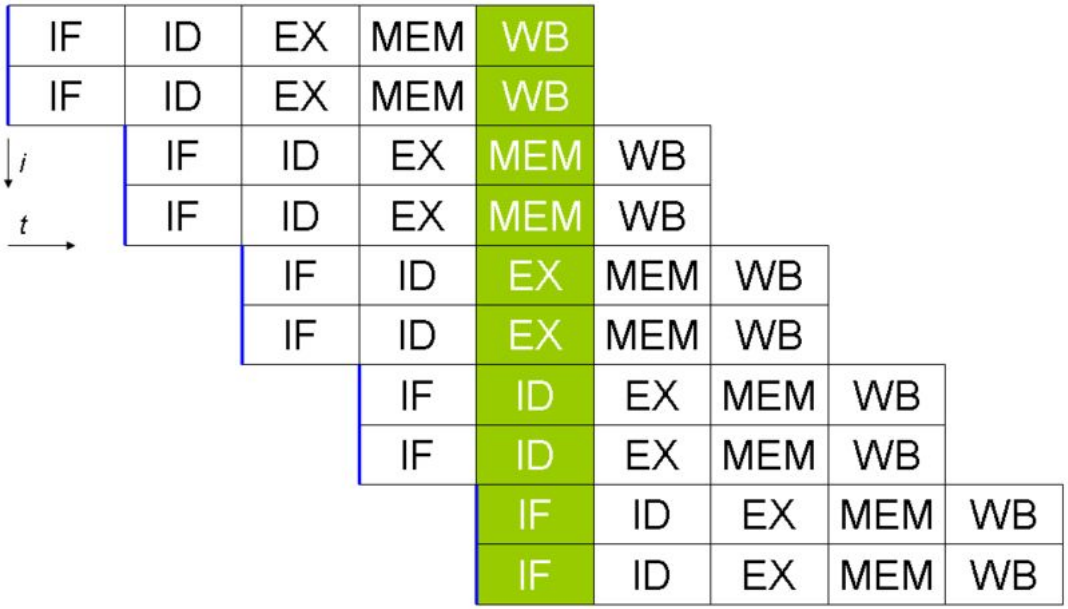
- NUMA, çok işlemcili makinelerde bellek erişim zamanının bellek yerine göre değiştiği bir bellek tasarımıdır.
- İlk defa 1990'larda ortaya çıkmıştır.
- Modern işlemciler, belleklere hızlı bir şekilde erişmeye ihtiyaç duyarlar. NUMA, istenen verinin "cache" bellekte bulunamaması, belleğin başka işlemci tarafından kullanılması gibi performans sorunlarını her işlemciye bellek vererek aşar.
- Intel Itanium ve AMD Opteron işlemciler ccNUMA tabanlıdır.





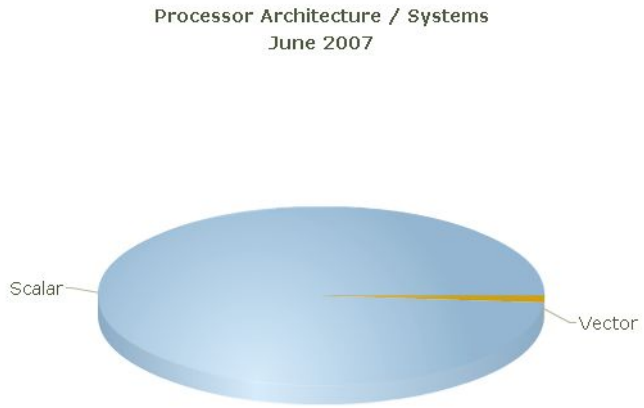
Superscalar İşlemciler

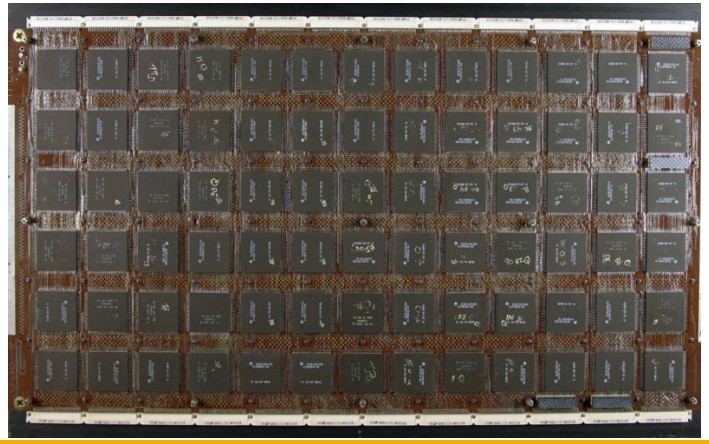
- 1998 senesinden beri üretilen bütün genel amaçlı işlemciler “superscalar” işlemcilerdir.
- “Superscalar” işlemci mimarisi, tek bir işlemcide makine kodu seviyesinde paralellik sağlar.
- “Superscalar” bir işlemci tek bir basamakta birden fazla işlem yapar.



Vektör Makineler

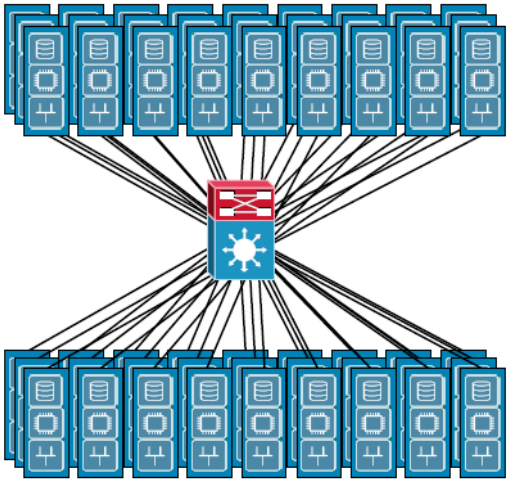
- Vektör işlemciler, aynı anda birden fazla veri üstünde matematik işlem yapabilen işlemcilerdir.
- Şu anda süperbilgisayar dünyasında vektör işlemciler çok az kullanılmaktadırlar.
- Ancak bugün çoğu işlemci vektör işleme komutları içermektedirler (Intel SSE).
- Vektör işlemciler, aynı matematiksel komutu farklı veriler üzerinde defalarca çalıştırmak yerine bütün veri yığınını alıp aynı işlemi yapabilirler.

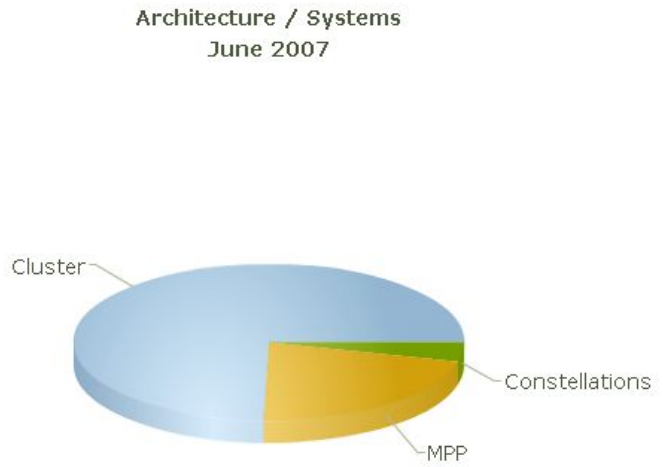




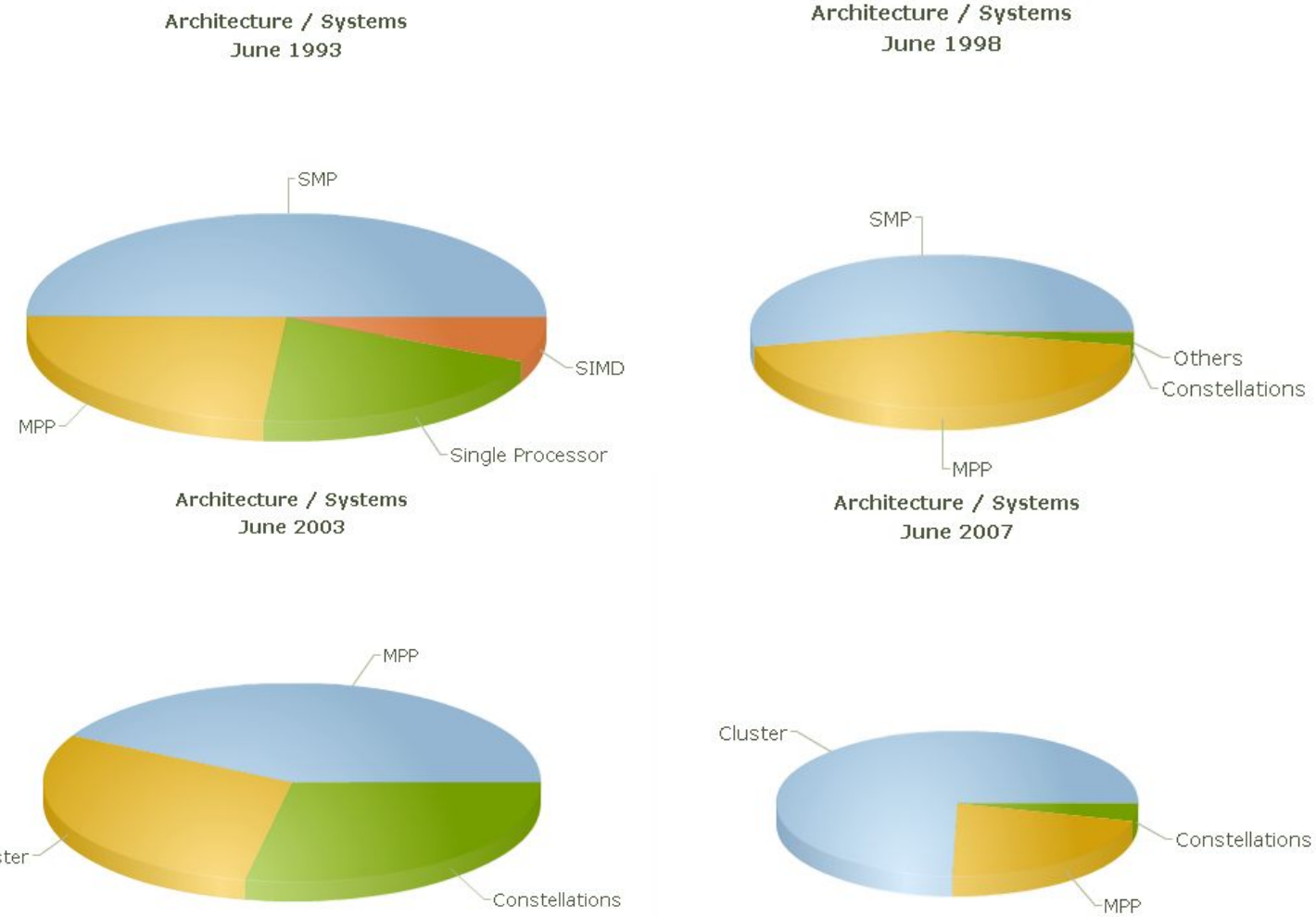
Küme Bilgisayarlar

- Hesaplamada küme bilgisayar kullanımı 1994 senesinde NASA'da Beowulf projesi ile başlamıştır. 16 Intel 486 DX4 işlemci ethernet ile bağlanmıştır.
- Yüksek performanslı hesaplama, artık küme bilgisayarlarla hesaplama halini almıştır.
- Küme bilgisayar, birlikte çalışmak üzere bağlanmış birden fazla sunucudan oluşur.
- En önemli dezavantajı kullanıcıya tek sistem arayüzü sunamamasıdır.



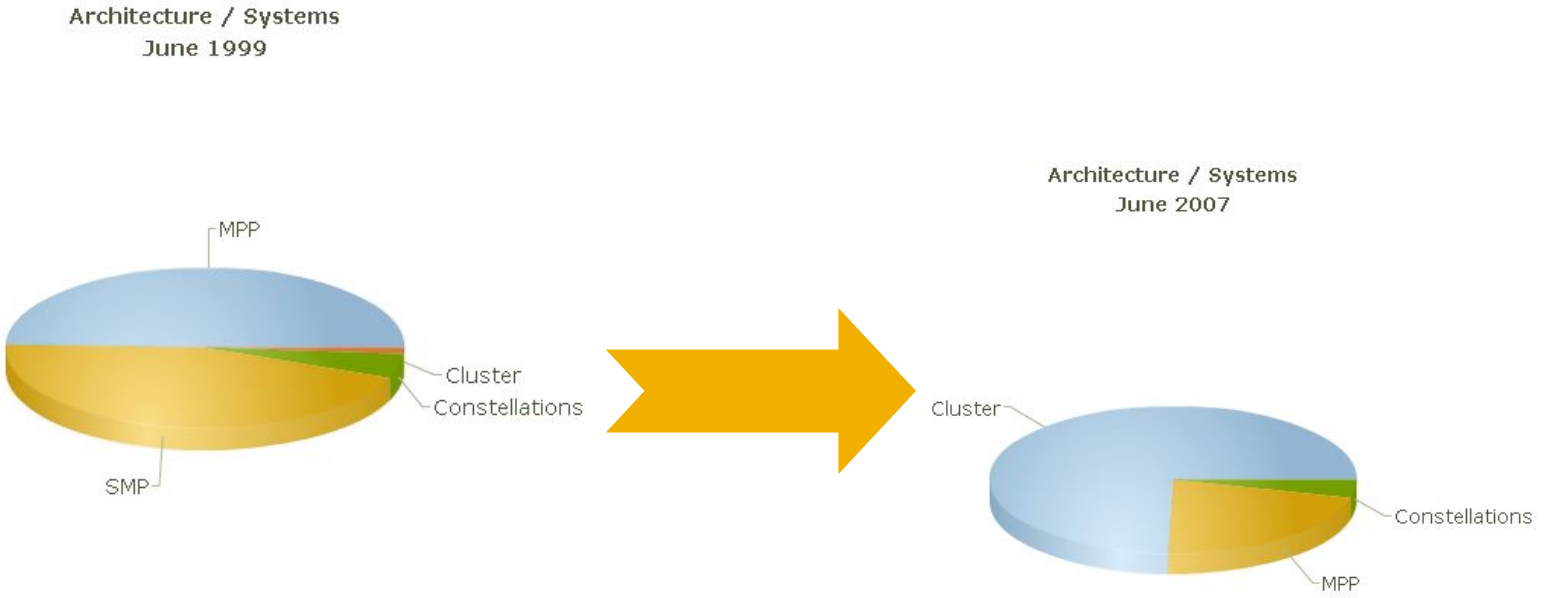


– TOP500 Listesine göre son 15 sene içinde süperbilgisayar sistemlerinde mimari değişimi



Neden kümeleme?

- Fiyat / performans
- Standardı oturmuş işletim sistemi, mesajlaşma gibi yazılım katmanları (Linux, MPI, OpenIB)
- Genişleyebilir, standardı oturmuş bağlantı teknolojileri (Gigabit Ethernet, Infiniband, 10 Gigabit Ethernet)
- Son senelerde süperbilgisayarların büyük bir kısmı küme bilgisayarlardan oluşmaktadır:



Kümeleme nedir?

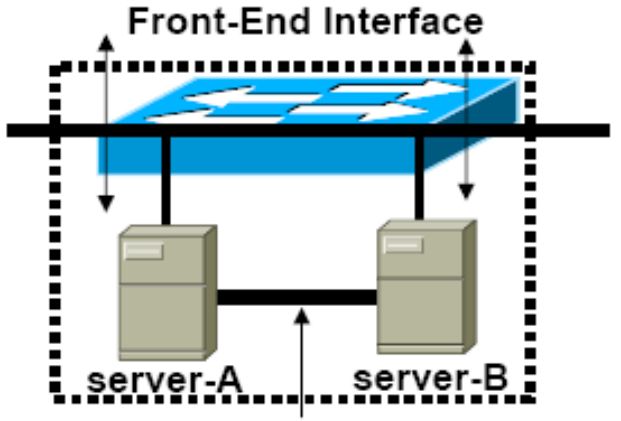
- Kümeleme iki veya daha fazla bilgisayarı:
  - Uygulama veya servis kullanılabilirliğini arttırmak için,
  - Yük dengelemek için,
  - Dağıtık ve yüksek başarımlı hesaplama için ağ ile birleştirmektir.

- Kümeleme değişik sistem katmanlarında gerçekleştirilebilir:
  - Depolama: Paylaşılmış disk, ikizlenmiş disk, paylaşılmayan veri
  - İşletim Sistemi: UNIX/Linux kümeleri, Microsoft (?) kümeleri
  - Uygulama Programlama Arayüzü: PVM, MPI
  - Uygulamalar

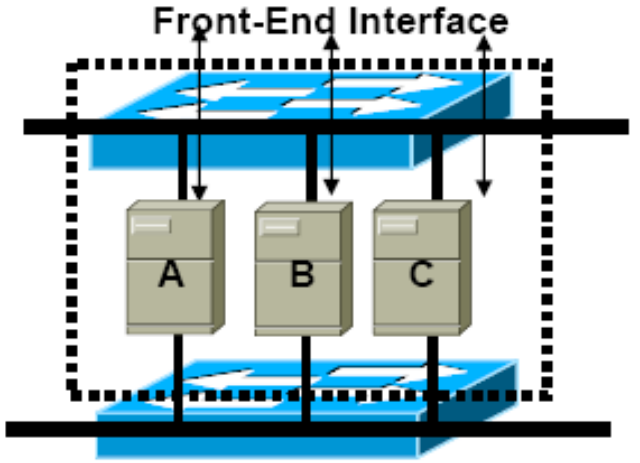
Dezavantajları

- Küme bilgisayarların önemli mimari dezavantajları vardır:
  - Ortak bellek yoktur.
  - İletişim bellek okuma/yazma hızına göre yavaştır.
- Bu kısıtlamalar uygulama için önemlidir. Uygulamanın bunlara göre de geliştirilmesi gerekebilir.
- Güç ve klima için genelde daha fazla miktarda bütçe gerekir.
- Ölçeklenebilirlik yakalamak bazı uygulamalar için zordur.

- HA kümeleri, servislerin ayakta kalma sürelerini arttırmak içindir.
- Aynı servisin birden fazla kopyası çevrimiçi veya çevrimdışı bekletilir. Serviste bir sorun olduğu zaman devreye alınır.
- Linux-HA projesi, sıklıkla bu amaçla kullanılan bir yazılımdır.

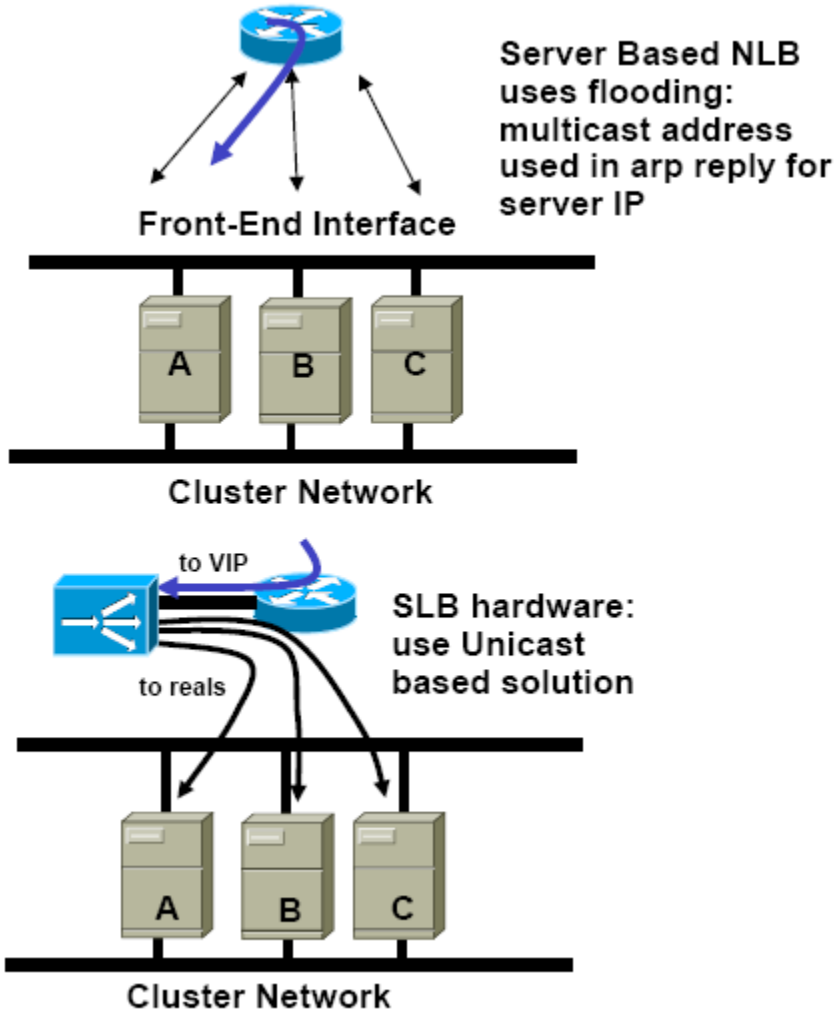


Cluster Network with 2 nodes -Crossover Cable Connection



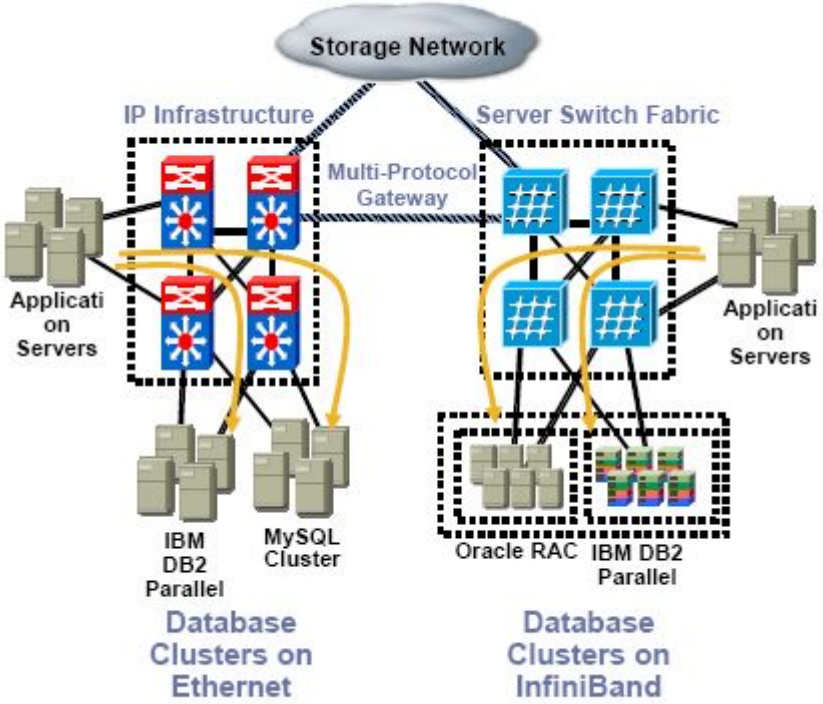
Cluster Network with 2 or more nodes on Switch VLAN

- Yük dengeleme kümeleri, ön arayüzden gelen bütün iş yükünü karşılayıp arkadaki sunuculara aktarırlar.
- Bu kümeler, sunucu çiftliği olarakta adlandırılırlar.
- LSF, MAUI, Sun Grid Engine gibi birçok yük dengeleyici yazılım vardır.
- “Linux Virtual Server” projesi de oldukça sık kullanılan bir yük dengeleyici çözümüdür.



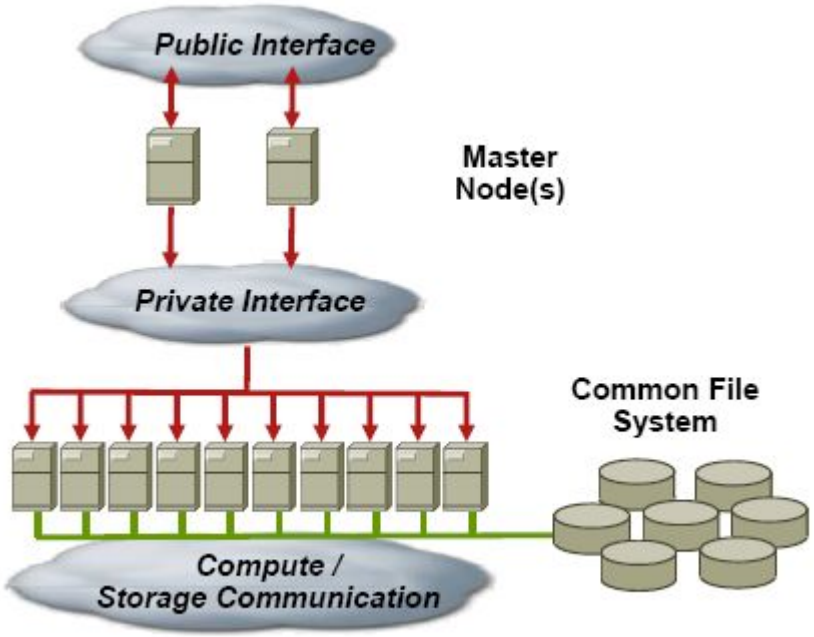
Veritabanı Kümeleri

- Son senelerde birçok veritabanı üreticisi, yüksek kullanılabilirlik, genişleyebilirlik ve yüksek başarımlılık için kümeleme teknolojisini için ürün çıkarmıştır.
- Bu çözümlerin bir kısmı paylaştırılmış disk alanı, bir kısmı ayrık veri alanları ile çözüm sunmaktadır.



YBH Kümeleri

- Bu kümeler, zaman kritik paralel, seri veya parametrik hesaplama işlerini çalıştırmak için kullanılır.
- Normal bir bilgisayarda inanılmaz sürede bitebilecek işlemci kritik uygulamaları çalıştırırlar.
- Genellikle normal PC veya sunucular ve Linux ile oluşturulan kümeler Beowulf ismini alırlar.
- MPI, YBH kümelerinde en çok tercih edilen haberleşme kütüphanesidir.



YBH Küme Mimarileri

- Yüksek başarımlı hesaplama ihtiyacını karşılamak isteyen bir kullanıcının önünde iki seçenek vardır:
  - Uygulamasına göre küme bilgisayarı edinmek.
  - Erişebildiği küme bilgisayarın özelliklerine göre uygulamasını geliştirmek, değiştirmek veya optimize etmek.

- Her iki durumda da bilinmesi veya hesaplanması gerekenler:
  - Uygulamanın özellikleri, gereksinimleri (yüksek bellek, her sunucuda yüksek miktarda geçici disk alanı, özel kütüphaneler ...),
  - Kümenin büyüklüğü (işlemci, bellek, disk),
  - Ağ bağlantı biçimi (gigabit ethernet, infiniband),
  - İşletim sistemi (Linux, Microsoft (?) ...),
  - Birçok kullanıcı veya grubun birlikte çalışabilirliği,
  - Derleyiciler (GNU, Intel, Portland Group ...)

Mimari


Kaynak İzleme/Yönetimi
PBS, MAUI, Condor, SGE
Derleyiciler
Kütüphaneler
Debuggerlar
MPI, PVM
Açık Kaynak Kod
Ticari
Sertifikalı
Güvenli
TCP/IP, GM, Infiniband
Ethernet, Gigabit Ethernet
Infiniband, Myrinet
Uygulamalar
Kuyruk Yöneticisi
Geliştirme Ortamları
Orta Katman/API
İşletim Sistemi
Bağlantı Protokolleri
Bağlantı Donanımı
Dosya Sistemi
Donanım/Yazılım
İzleme/Yönetim
Yükleme
Çalışma Ortamı
PVFS, Ticari Sistemler
Fiber Kanal, SCSI
Dahili, Harici
NFS
IPMI, ACPI, SNMP, HFS
Donanım Testleri
Ganglia, MRTG
IPMI, ACPI, PXE
İmaj Sunucuları
FAI, Kickstart
Soğutma
Güç Kaynakları

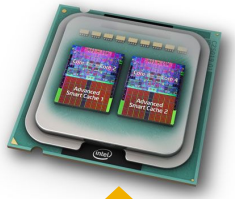
Sunucular

- Günümüzde 1U boyutta 16 çekirdekli sunucular almak mümkün olmaktadır.
- Küme bilgisayarlarda sunucu seçimi konusunda birçok faktör vardır:
  - İşlemciler : Tek çekirdek, çok çekirdek, çoklu işlemci soketi ...
  - Anakart : PCI-X, PCI-Express, HyperTransport ...
  - Sunucu form faktörü : Blade, rack monte, PC ...
  - Bellek : Boyutu, DDR-2, DDR-3, FBDIMM …
  - Disk : Boyutu, SATA, SCSI, SAS …
  - Ağ bileşenleri : Gigabit Ethernet, Infiniband, Quadrics ...

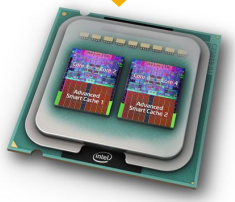
Bileşenlerin bant genişliği

Infiniband DDR 4x
2 GB/s

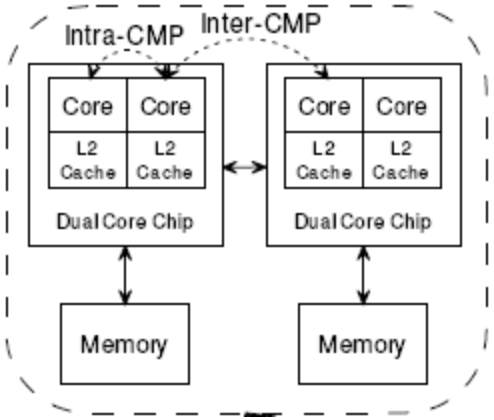
SATA-300
300 MB/s



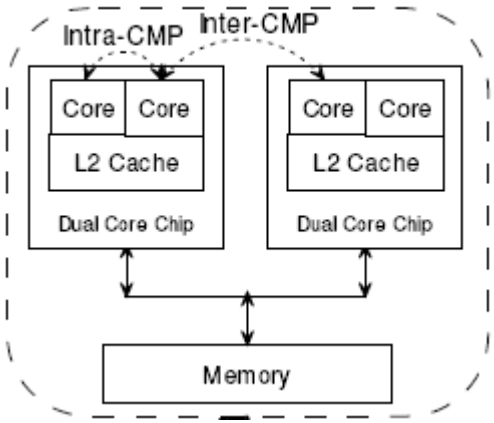
PC2-5300
5.3 GB/s



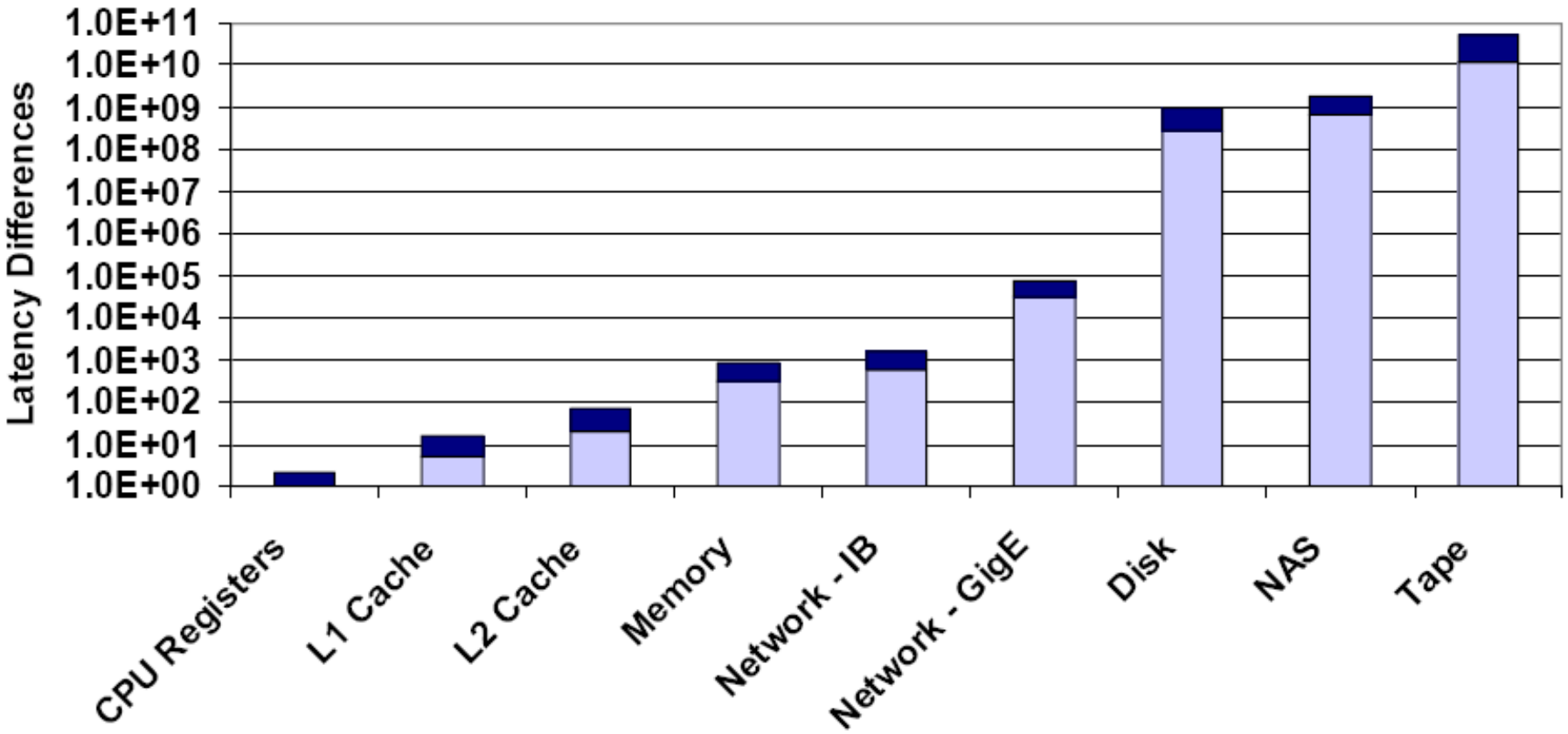
NUMA-based Design



Bus-based Design

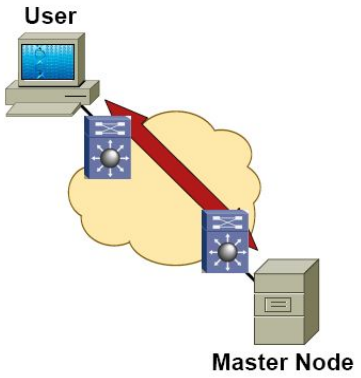


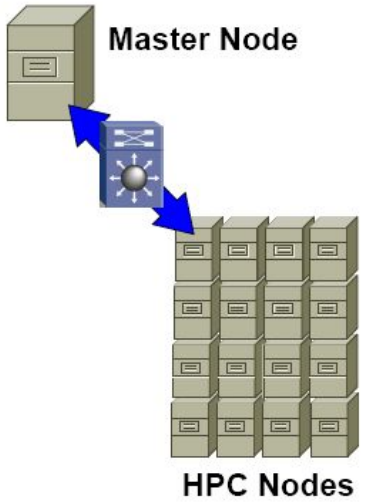
Gecikme Süreleri



Ağ Bağlantısı

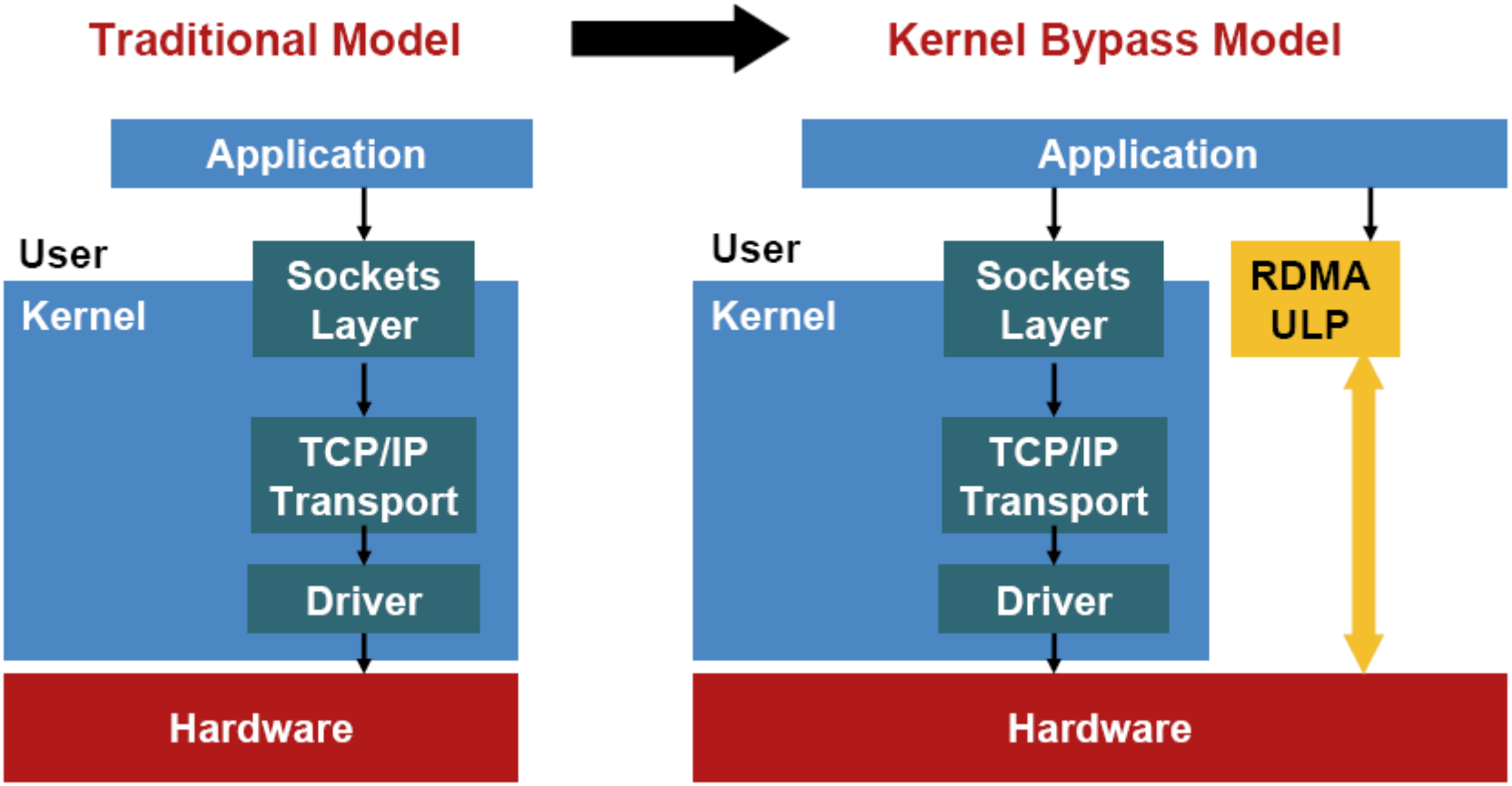
- Genellikle tek bir kümede birden fazla ağ bulunur:
  - Kullanıcı ağı:
    - *İş göndermek, görselleştirme, sonuç görüntüleme için kullanılır.*
    - *Grid haberleşmesi için de kullanılabilir.*
    - *Kümelere bağlanmak için genellikle ssh kullanılır.*
  - Yönetim ağı:
    - *İş planlamak, sunucuları izlemek, kurmak için kullanılır.*
    - *Genellikle IP üzerinden çalışırlar.*
    - *Ganglia gibi yazılımlar multicast çalışırlar.*





IPC Ağı

- Küme bilgisayar performansı ve verimi IPC ağı tarafından belirlenir. Haberleşmede harcanan her fazla süre daha az işlem zamanı demektir.
- Günümüzde küçük kümeler ve gevşek bağlı uygulamalar için gigabit ethernet ideal bir çözümdür.
- Büyük kümeler ve sıkı bağlı uygulamalar için Infiniband, Quadrics gibi çözümler vardır.
- Uygulama gereksinimlerini anlamak teknoloji seçiminde çok önemlidir.

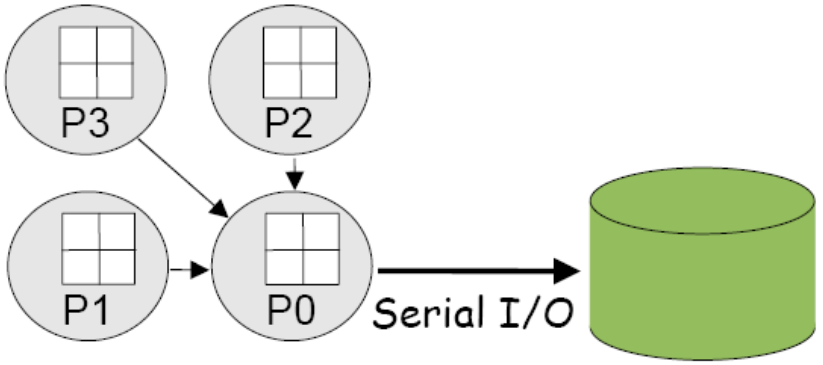

Traditional Model
Kernel Bypass Model
Application
User
Kernel
Sockets Layer
TCP/IP Transport
Driver
Hardware
Application
User
Kernel
Sockets Layer
RDMA ULP
TCP/IP Transport
Driver
Hardware

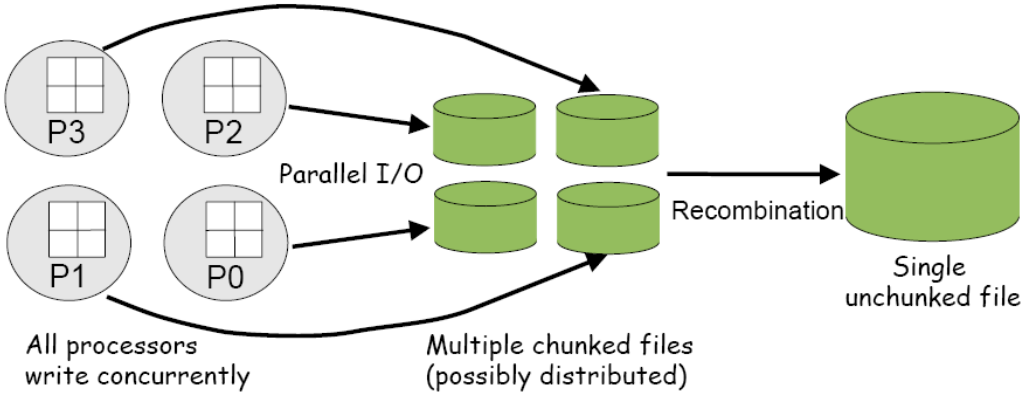
Depolama

- Küme bilgisayarlarda çalışan kullanıcıların farklı depolama ihtiyaçları bulunur.
  - Ev dizini, uygulamalar için ortak veri alanı
  - Yığın veri saklamak için veri ambarları
  - Yedekleme ve yığın veriler için tape üniteleri
  - Bazı uygulamalar için sunucularda geçici paylaşılmayan disk alanları
- Küme bilgisayarlarda hesaplama yapılan sunucularda kurulum diski veya geçici disk alanı bulundurmak gerekli değildir. Ancak çoğu durumda maliyeti düşüren bu çözüm tercih edilmemektedir.
- Uygulama performansı için özellikle paylaşılan disk alanlarının ihtiyaca uygun tasarlanması gerekir.

Seri ve Paralel G/Ç

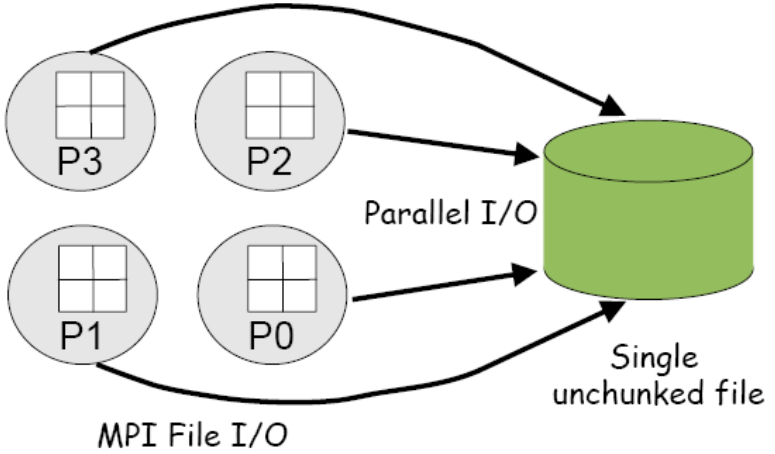
- P0 bütün işlemcilerden veriyi toplar.
- G/Ç işlemcisi darboğazdır.





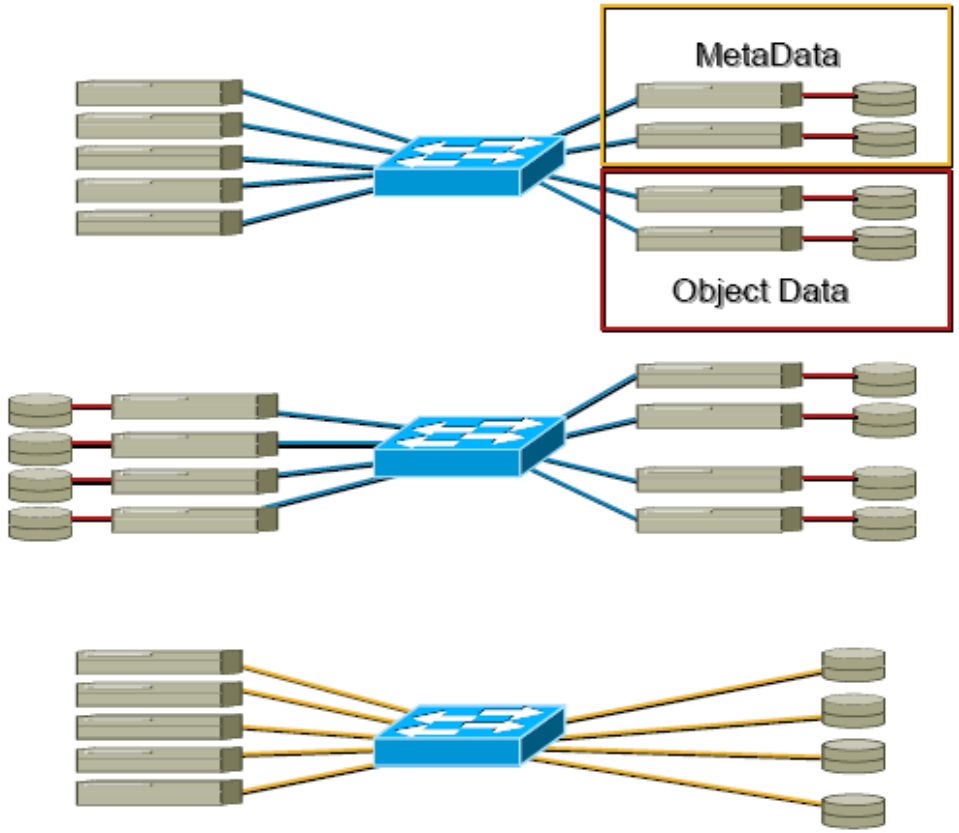
- Sistemde açık dosya sayısı limitlidir.
- Parçalanmış dosyalar birleştirilmelidir.

- MPI G/Ç, paralel dosya sistemi ihtiyacı duyar.

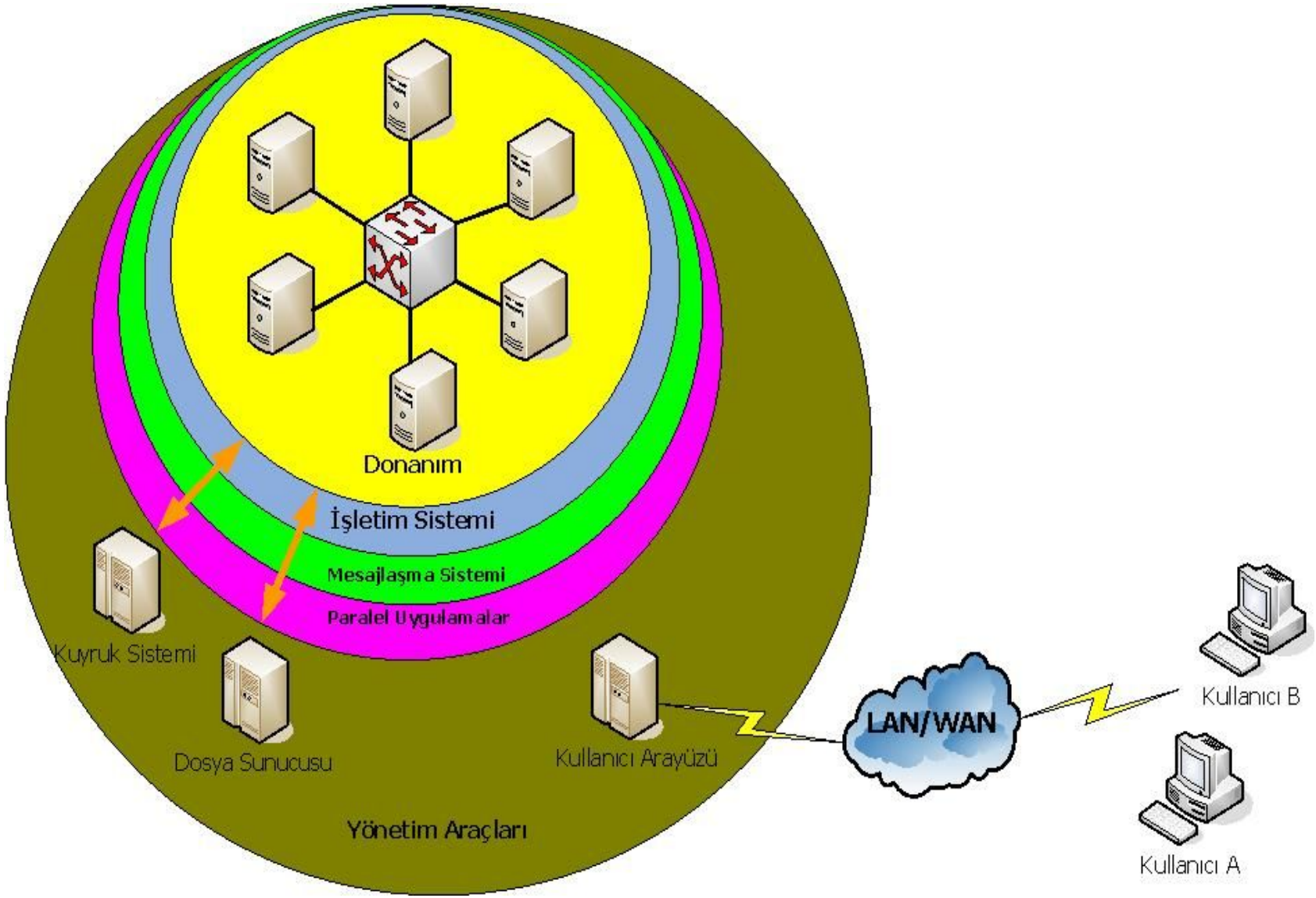


Dosya Sistemleri

- Paralel olmayanlar:
  - NFS, CIFS
- Paralel (“Metadata”)
  - Lustre : Ölçeklenebilir
  - Panasas : Ölçeklenebilir
- Paralel (“Metadata” olmadan)
  - XFS
  - IBM GPFS : Ölçeklenebilir
  - PVFS
  - Oracle Cluster FS



Mimari - Yazılım



Küme Bilgisayarlar - Örnek: Deniz