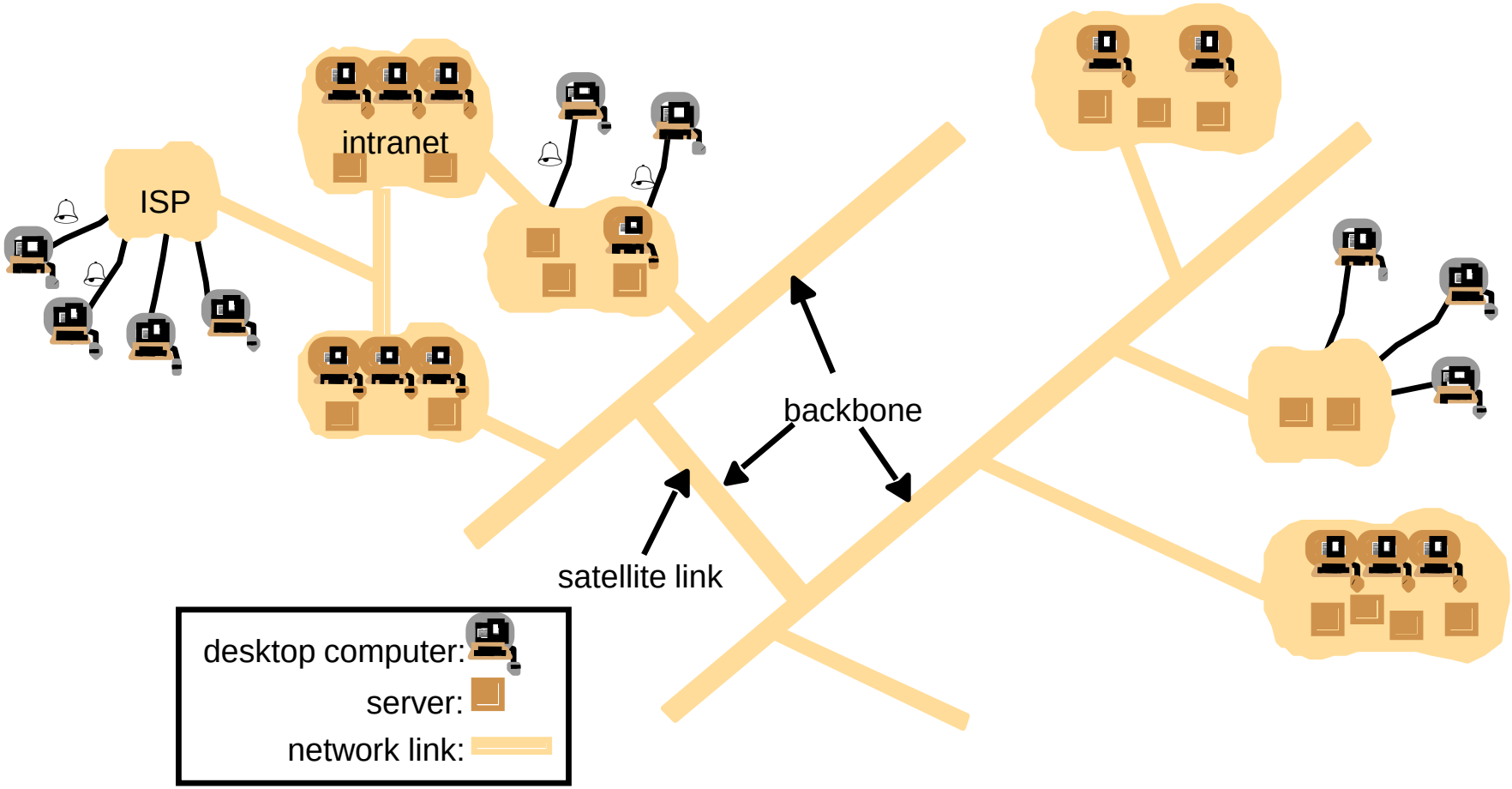
Distributed Systems

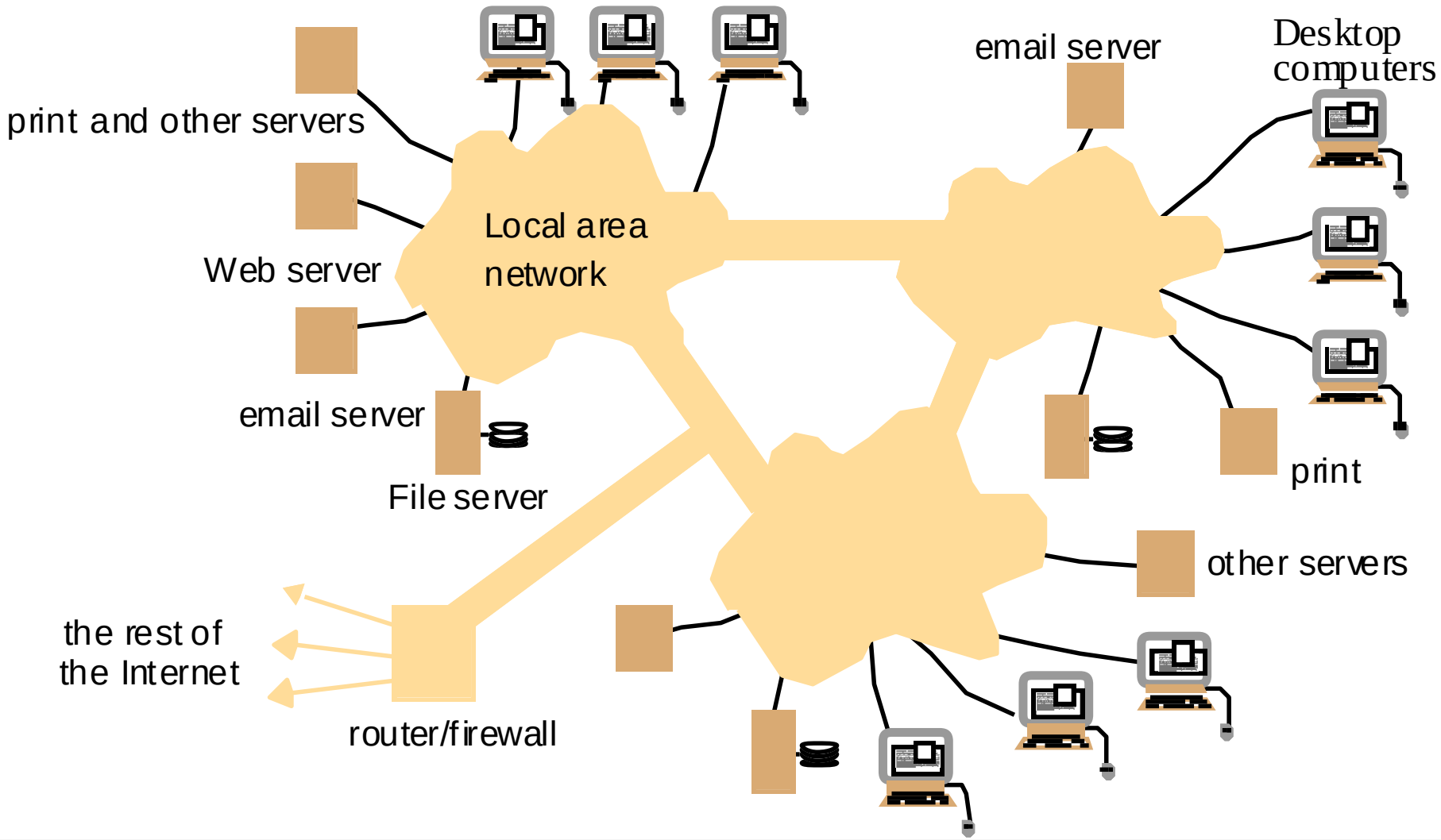
Current State of Computing

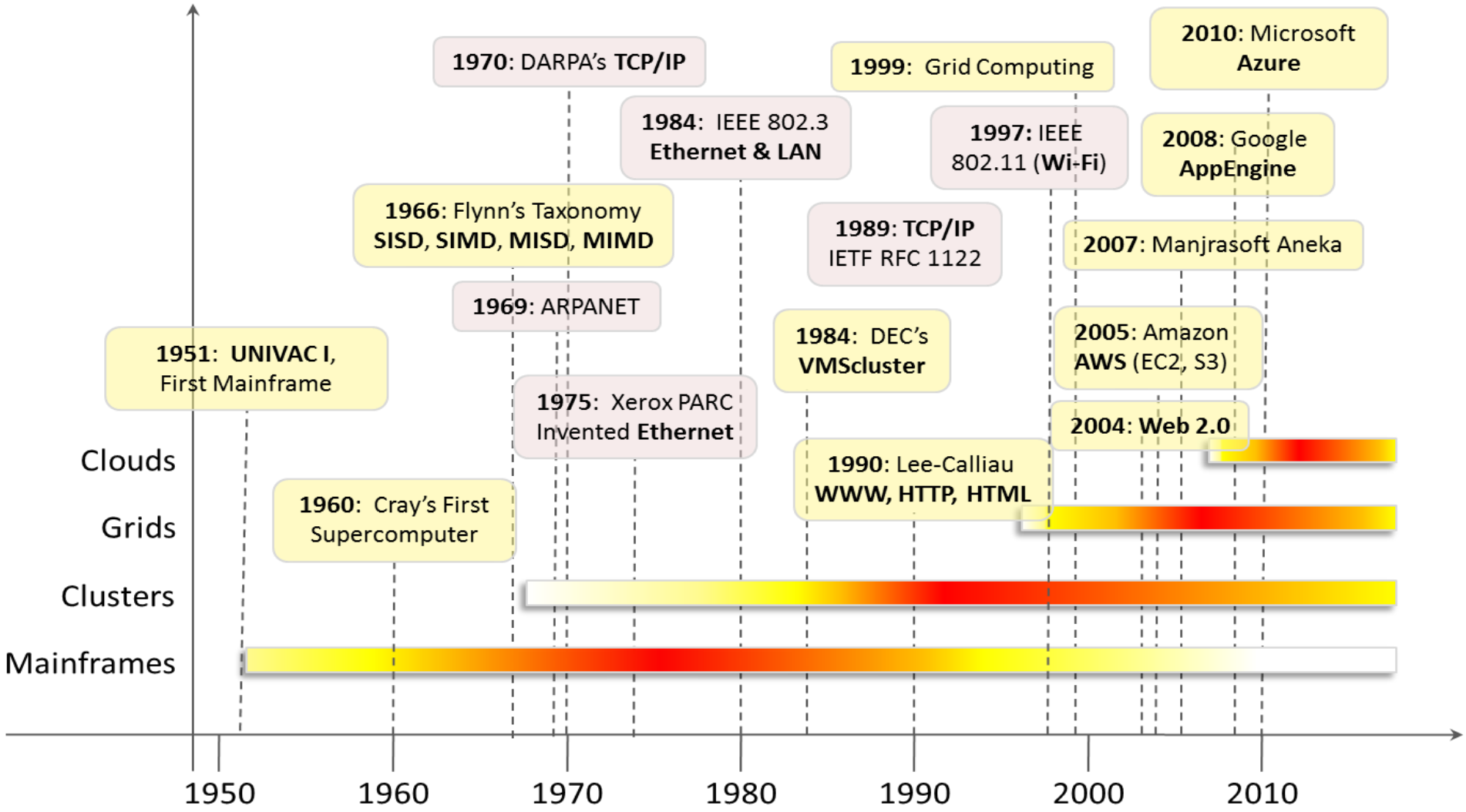
Characterisation of Distributed Systems

A typical portion of the Internet

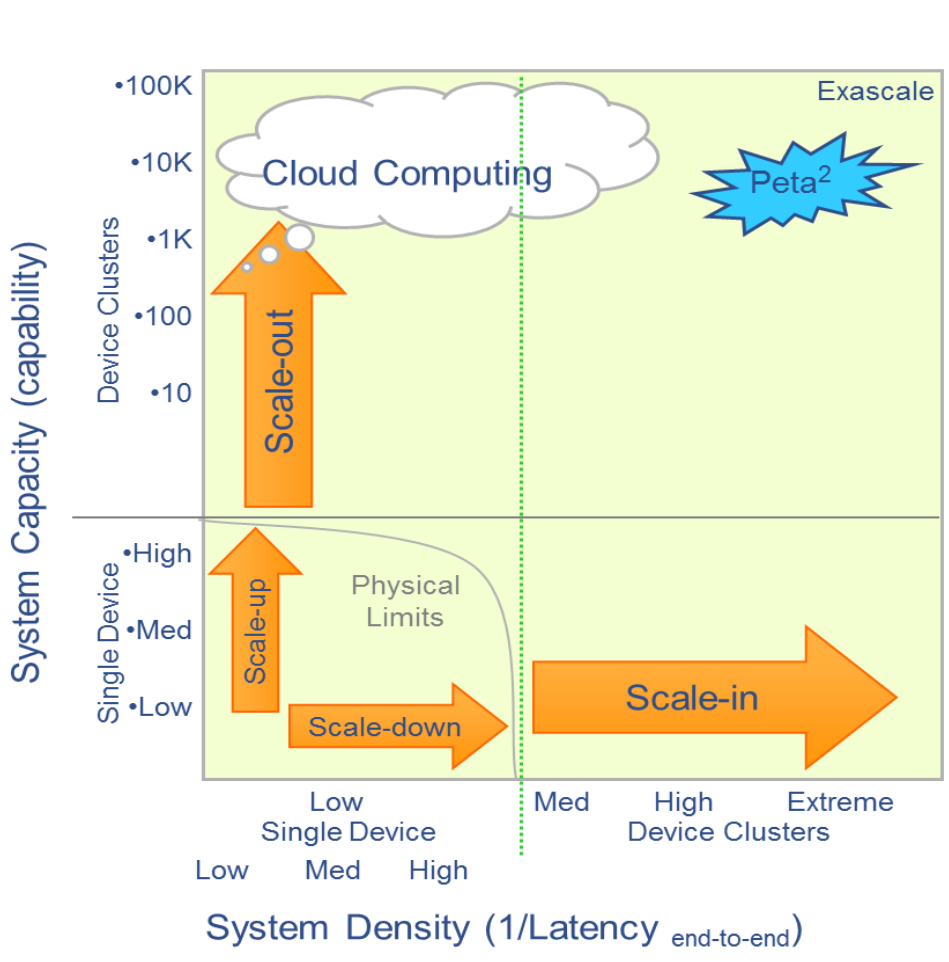


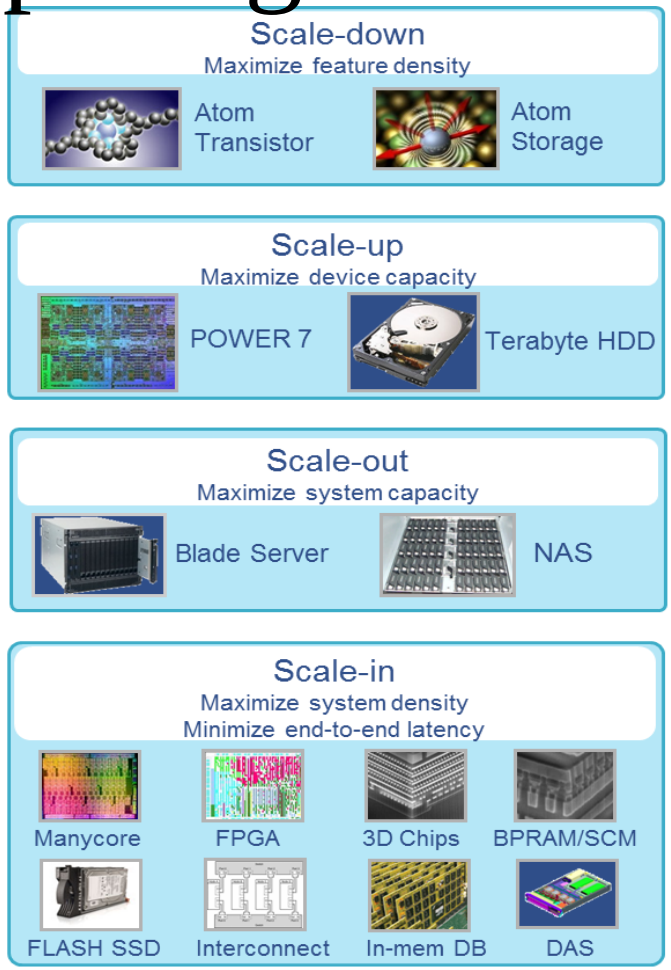
A typical intranet




1970: DARPA's TCP/IP
1999: Grid Computing
2010: Microsoft Azure
1984: IEEE 802.3 Ethernet & LAN
1997: IEEE 802.11 (Wi-Fi)
2008: Google AppEngine
1966: Flynn's Taxonomy SISD, SIMD, MISD, MIMD
1989: TCP/IP IETF RFC 1122
2007: Manjrasoft Aneka
1969: ARPANET
1984: DEC's VMScluster
2005: Amazon AWS (EC2, S3)
1951: UNIVAC I, First Mainframe
1975: Xerox PARC Invented Ethernet
2004: Web 2.0
1990: Lee-Calliau WWW, HTTP, HTML
1960: Cray's First Supercomputer
Clouds
Grids
Clusters
Mainframes
1950
1960
1970
1980
1990
2000
2010

Trends in Computing





Source: IBM

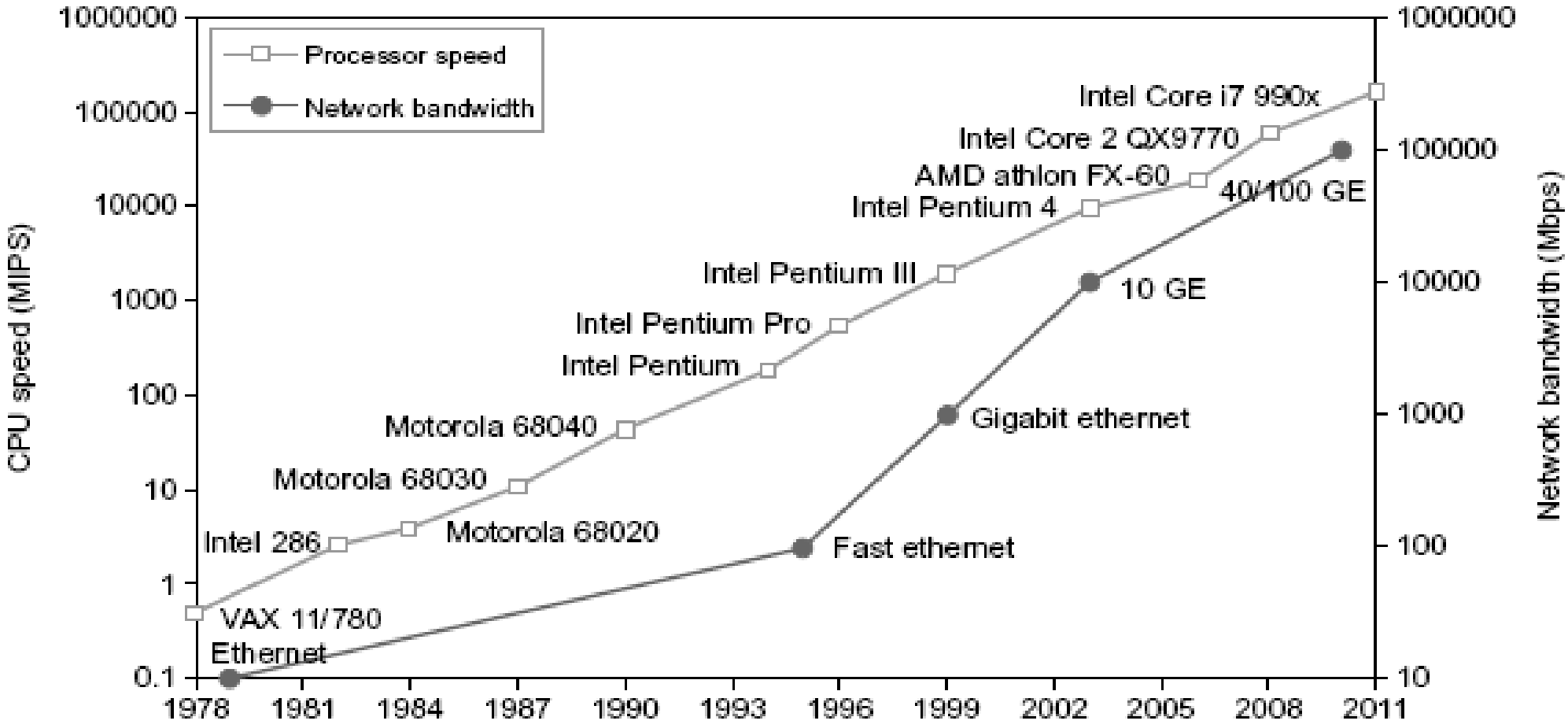


FIGURE 1.4

Improvement in processor and network technologies over 33 years.

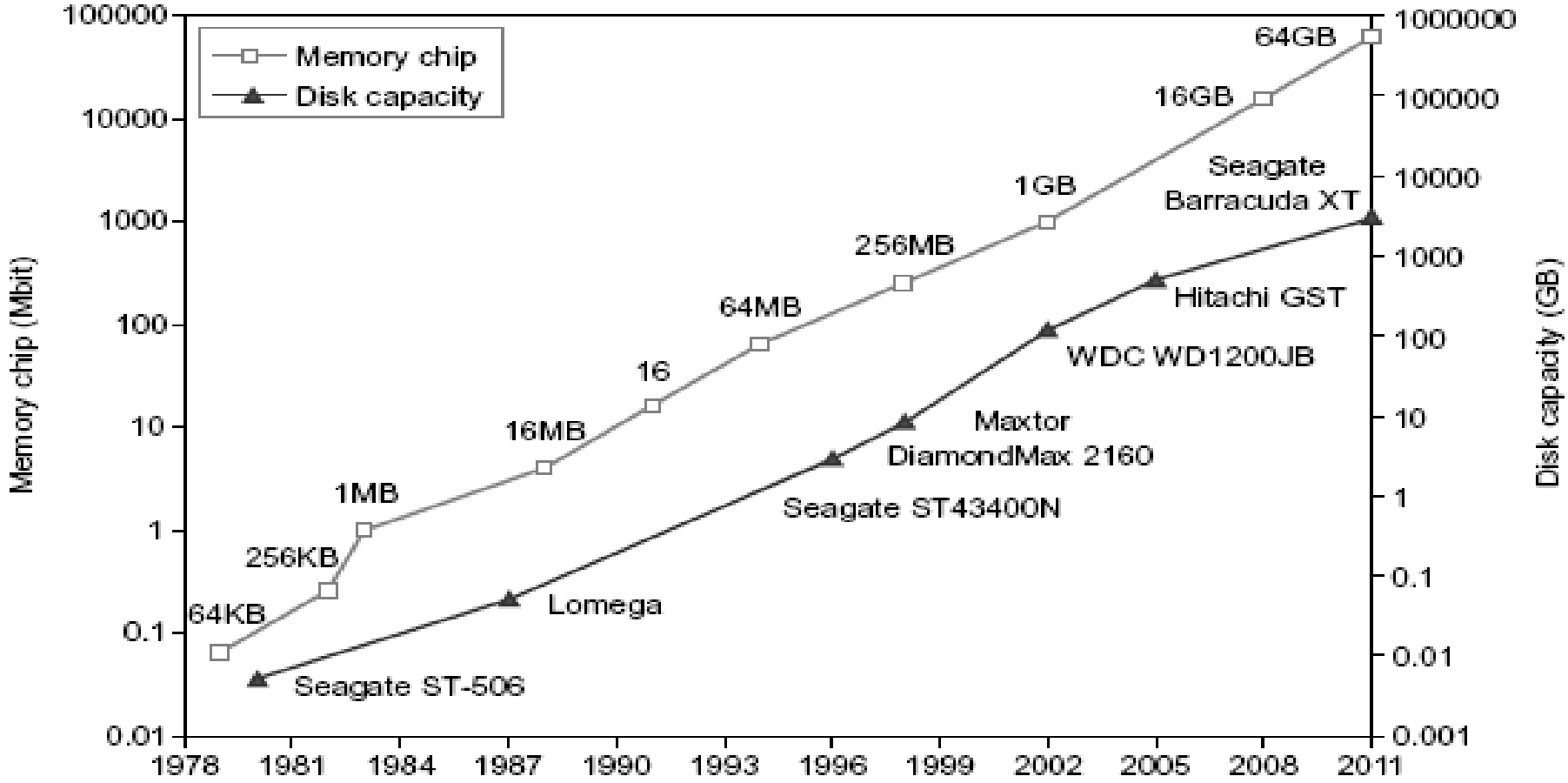


FIGURE 1.10

Improvement in memory and disk technologies over 33 years. The Seagate Barracuda XT disk has a capacity of 3 TB in 2011.

(Courtesy of Xiaosong Lou and Lizhong Chen of University of Southern California, 2011)

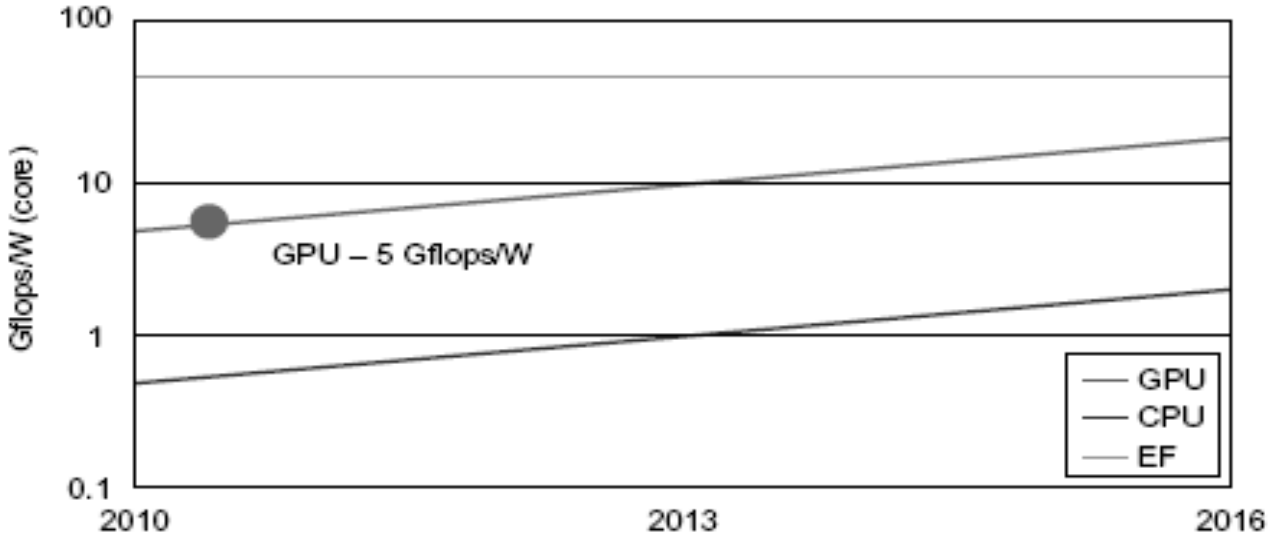


FIGURE 1.9

GPU and CPU performance in Gflops/Watt/core, compared with 60 Gflops/Watt/core projected in future Exascale systems.

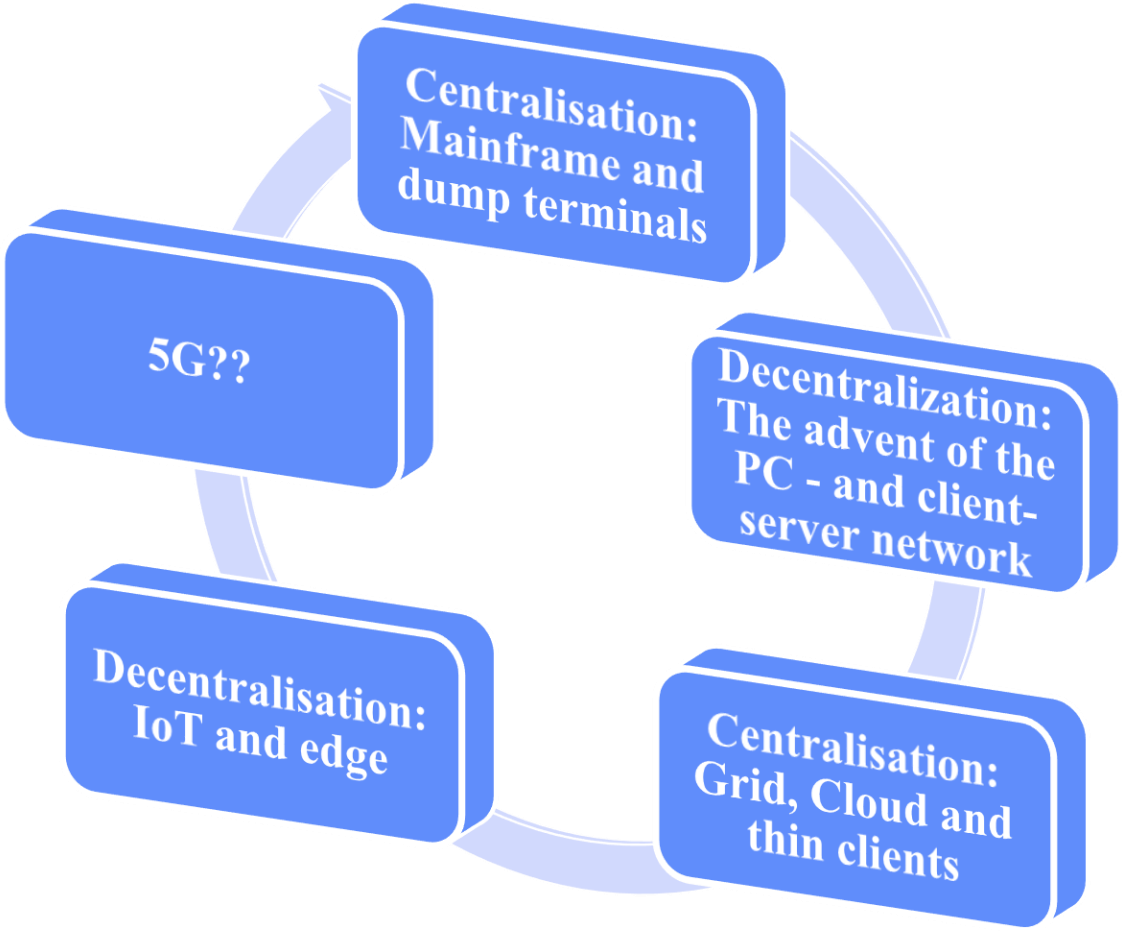
Scalability: Computers in the Internet

Date	*Computers*	*Web servers*
1979, Dec.	188	0
1989, July	130,000	0
1999, July	56,218,000	5,560,866
2003, Jan.	171,638,297	35,424,956

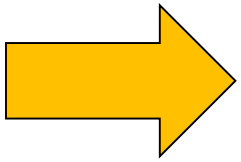
Computers vs. Web servers in the Internet

Date	*Computers*	*Web servers*	*Percentage*
1993, July	1,776,000	130	0.008
1995, July	6,642,000	23,500	0.4
1997, July	19,540,000	1,203,096	6
1999, July	56,218,000	6,598,697	12
2001, July	125,888,197	31,299,592	25
		42,298,371	

What happens next
Distributed Systems Cycle

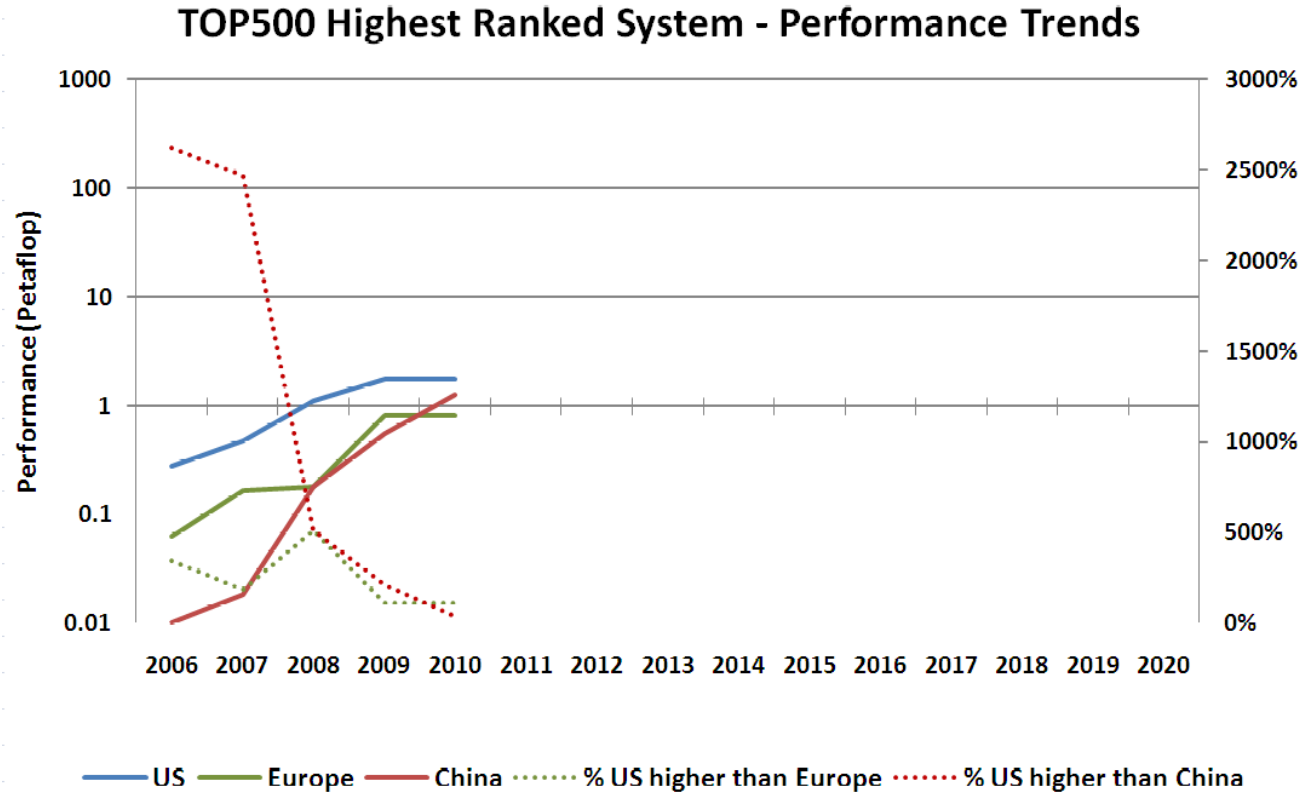


- Petascale Computing (10^{15})
- Multicore computing
  - 1-24 cores commodity architectures
  - 100+cores proprietary architectures
  - 400+ GPU cores

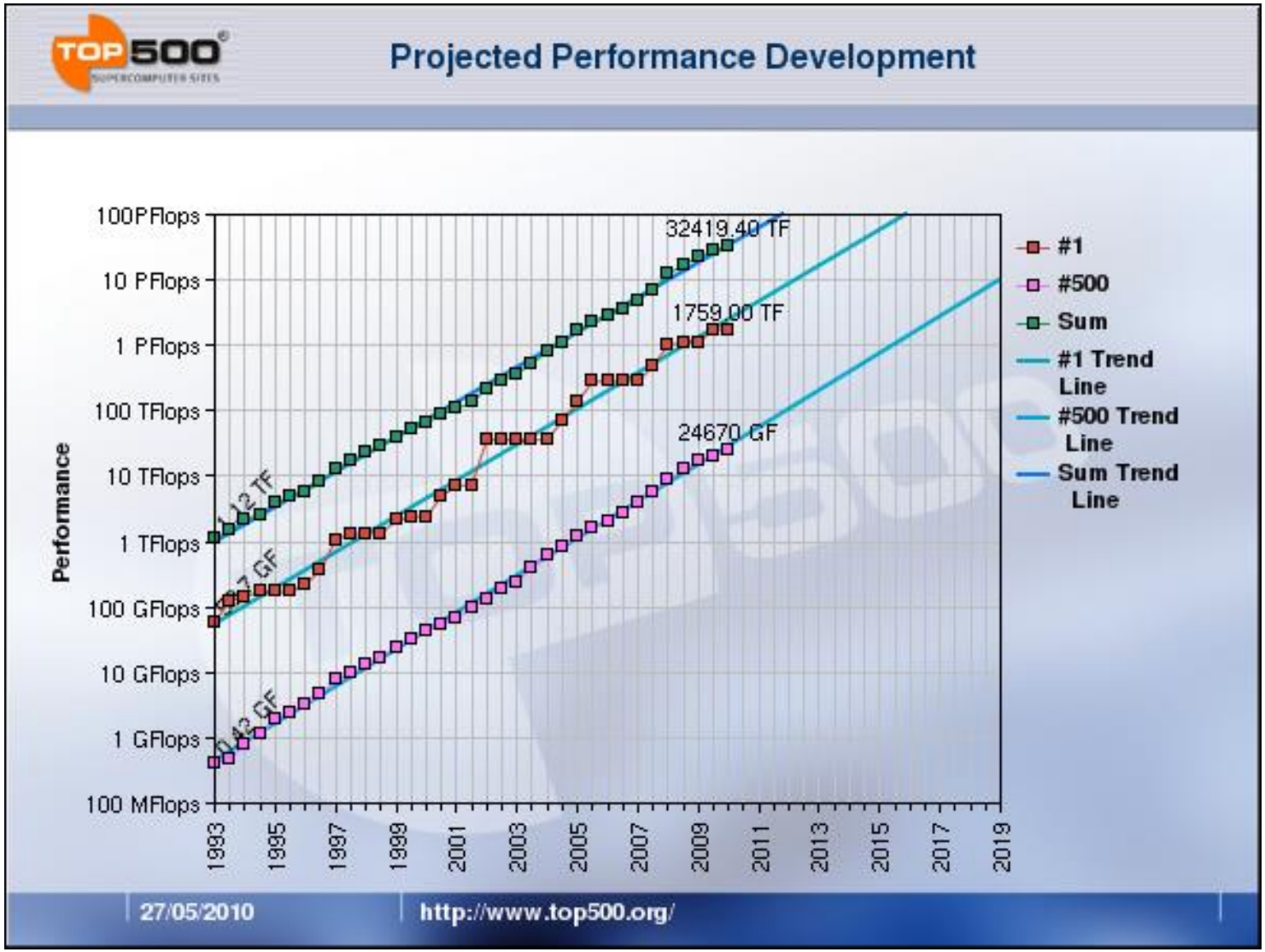


- Exascale Computing (10^{18})
- Manycore computing
  - ~1000-core commodity architectures (heterogeneous, merged with GPUs etc)
- 1M nodes
- 1B processor cores

Towards Exascale Computing



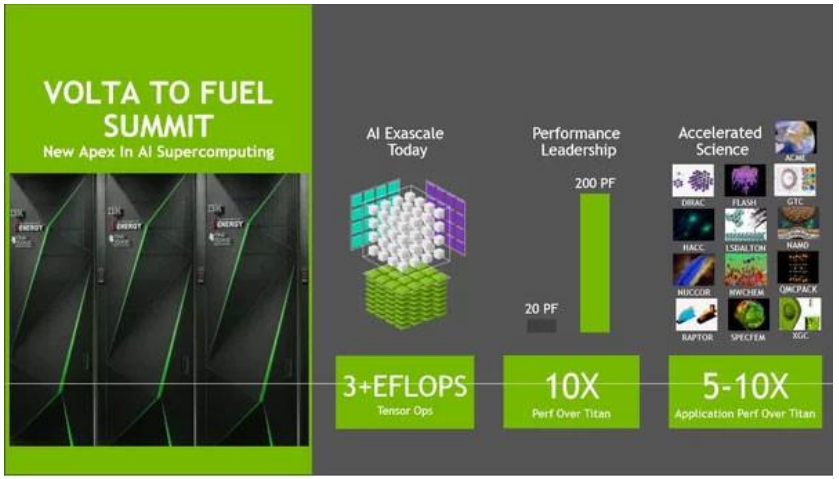
- In 2006, the performance leadership of the US's highest performance system was more than 2500% higher than the best system in China, in the first half of 2010, the gap decreased to less than 40% (**Source: US HPC Advisory Council**)

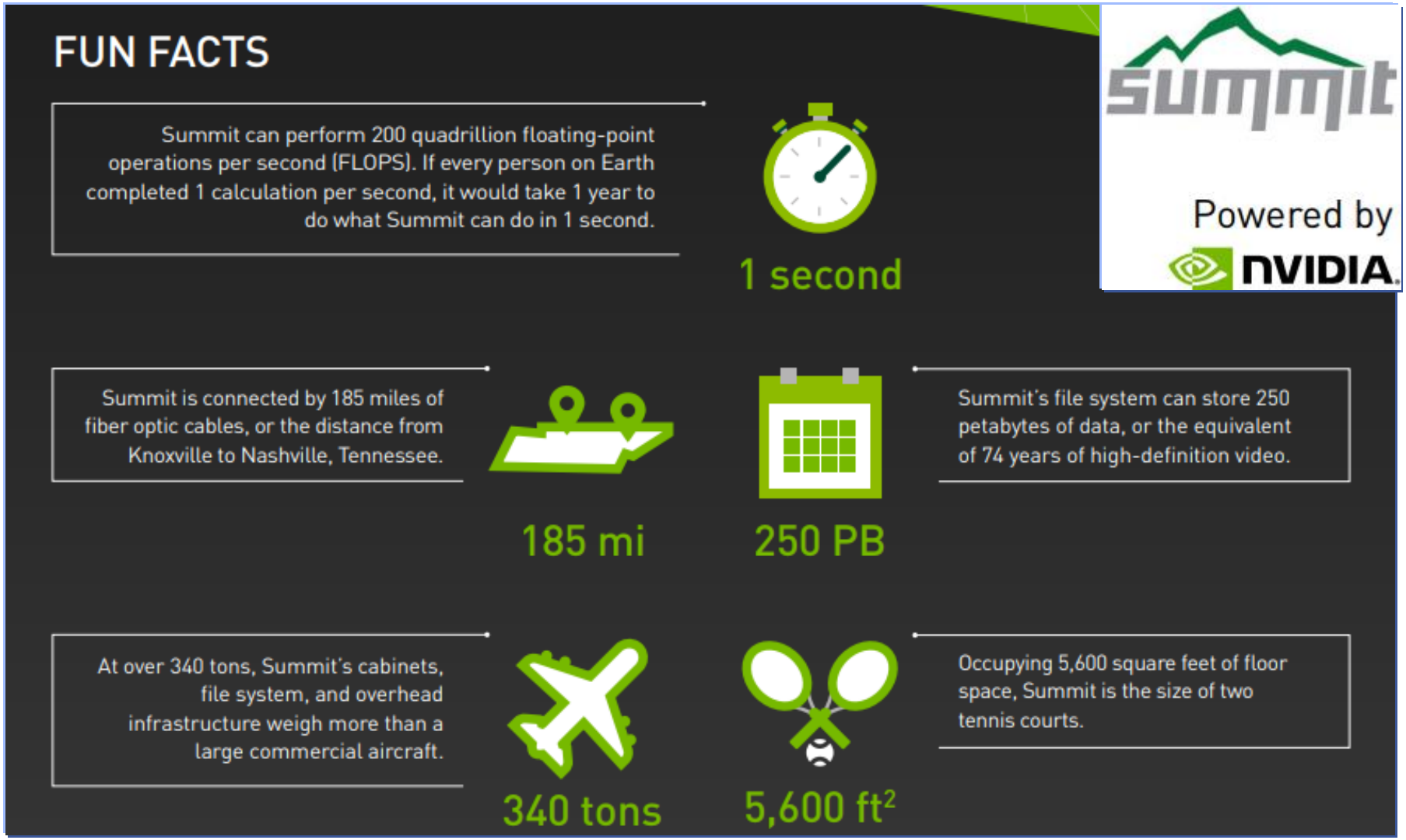

TOP500
SUPERCOMPUTER SITES
Projected Performance Development
100 PFlops
10 PFlops
1 PFlops
100 TFlops
10 TFlops
1 TFlops
100 GFlops
10 GFlops
1 GFlops
100 MFlops
Performance
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
32419.40 TF
1759.00 TF
24670 GF
1.12 TF
#1
#500
Sum
#1 Trend Line
#500 Trend Line
Sum Trend Line
27/05/2010
http://www.top500.org/

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,397,824	143,500.0	200,794.9	9,783
2	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
3	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	387,872	21,230.0	27,154.3	2,384
6	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/NNSA/LANL/SNL United States	979,072	20,158.7	41,461.2	7,578
7	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	32,576.6	1,649
8	**SuperMUC-NG** - ThinkSystem SD530, Xeon Platinum 8174 24C 3.1GHz, Intel Omni-Path , Lenovo Leibniz Rechenzentrum Germany	305,856	19,476.6	26,873.9	
9	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209
10	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , IBM DOE/NNSA/LLNL United States	1,572,864	17,173.2	20,132.7	7,890

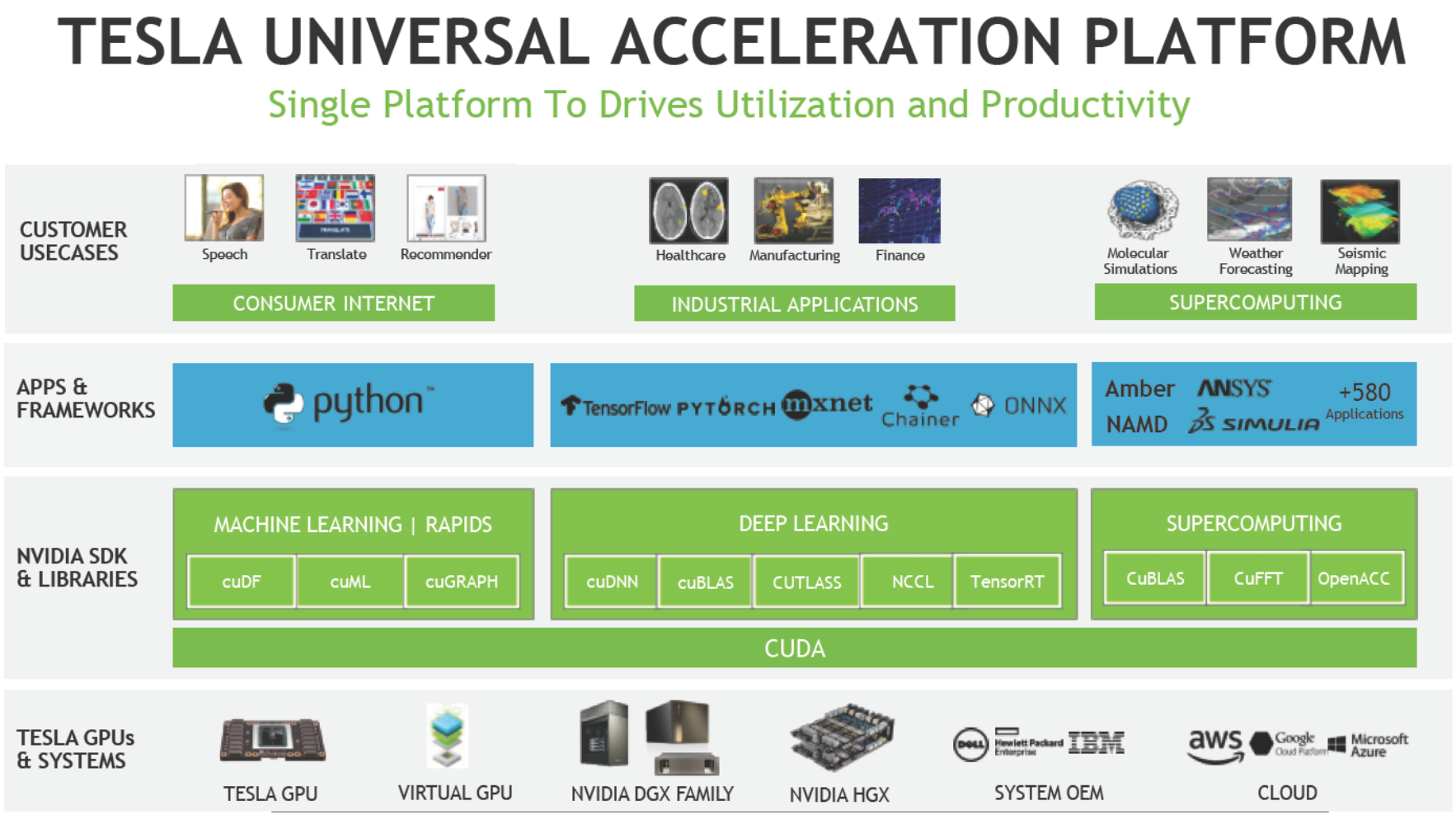
- Oak Ridge National Laboratory in Tennessee
- **9,216 IBM Power9** processor chips running at 3.1GHz, and each of those has 22 processing cores
- Connected to each pair of Power9 chips are six Nvidia Volta Tensor V100 graphics chips
- **27,648 V100**
- 200 PetaFLOPS

- 4,608 interconnected computer nodes housed in refrigerator-sized cabinets
- liquid-cooled by pumping 4,000 gallons of water per minute through the system.




FUN FACTS
Summit can perform 200 quadrillion floating-point operations per second (FLOPS). If every person on Earth completed 1 calculation per second, it would take 1 year to do what Summit can do in 1 second.
1 second
summit
Powered by
NVIDIA
Summit is connected by 185 miles of fiber optic cables, or the distance from Knoxville to Nashville, Tennessee.
185 mi
250 PB
Summit's file system can store 250 petabytes of data, or the equivalent of 74 years of high-definition video.
At over 340 tons, Summit's cabinets, file system, and overhead infrastructure weigh more than a large commercial aircraft.
340 tons
5,600 ft²
Occupying 5,600 square feet of floor space, Summit is the size of two tennis courts.

NVIDIA Universal Acceleration Platform for HPC and AI



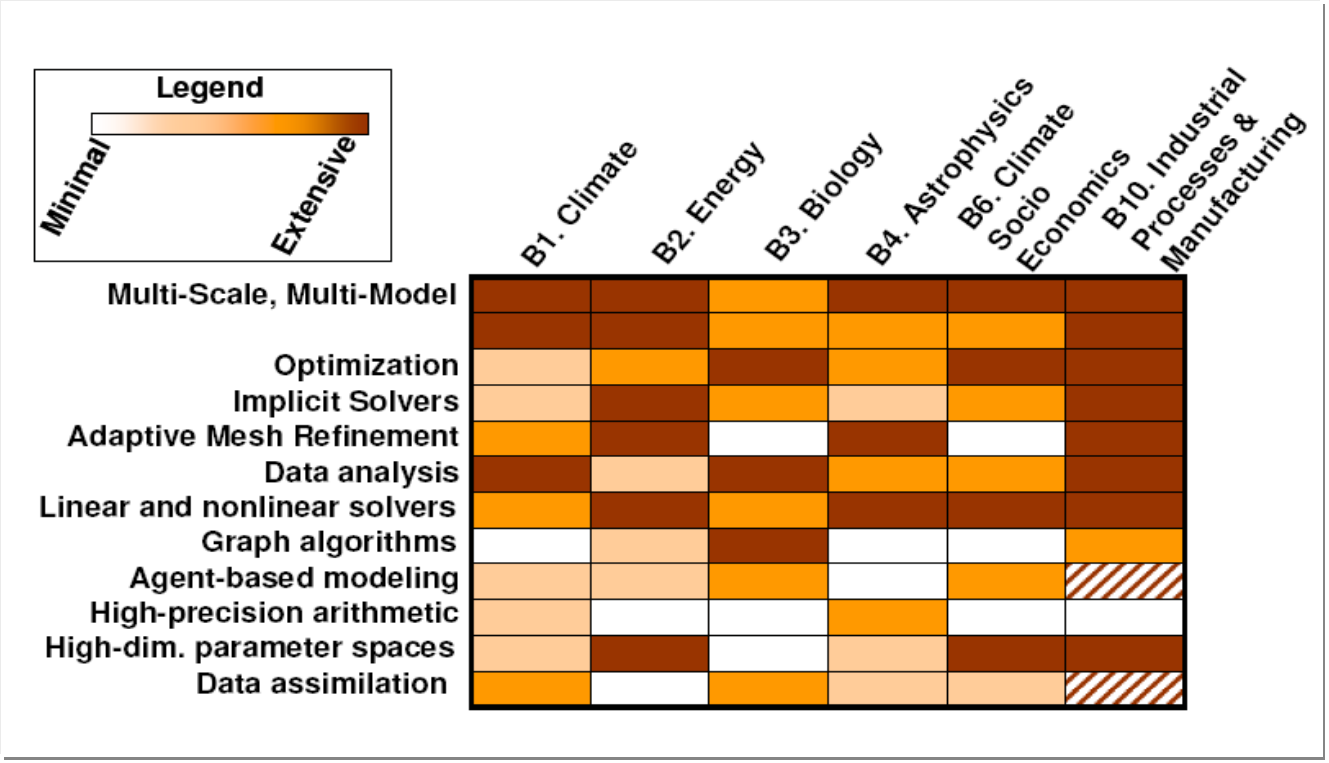
Source NVIDIA, 2018

- One of the world's top-10 AI supercomputers
- A cluster supercomputer with 660 NVIDIA DGX-1 nodes.
- Each such node packs eight NVIDIA GV100 GPUs
- 5,280 GPUs total

Sunway TaihuLight - 神威·太湖之光

- 40,960 Chinese-designed Sunway SW26010 manycore 64-bit RISC processors based on the architecture
- Each processor chip contains 256 processing cores, and an additional four auxiliary cores for system management
- A total of 10,649,600 CPU cores across the entire system
- 93 PetaFLOPS

Trends in Computing



Source: ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems, DARPA IPTO

Trends in Computing

- Data volume, velocity, and variety is growing at an astounding rate with a full 90% of the world's data less than two years old.
- "Big Data is big. It's 2.5 quintillion bytes of data every day big."
- Almost 90% of this data is unstructured

Homeland Security

•600,000 records/sec

Telco Promotions

•100,000 records/sec

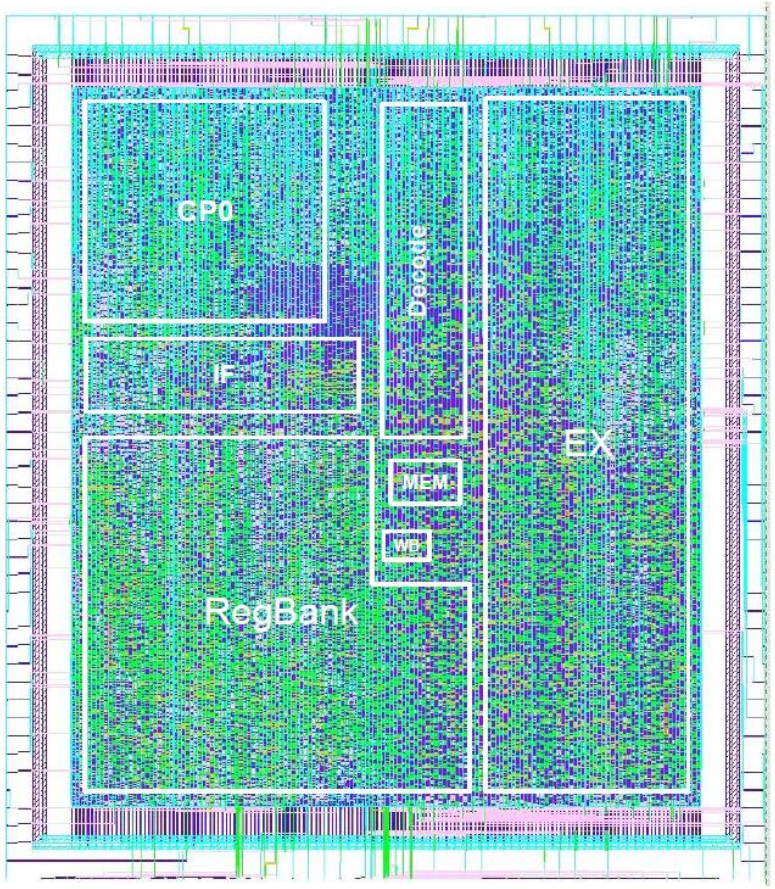
- 300M users
- 10000 data centres
- 30000 servers
- 25 Terabytes of Log Data - daily

Traffic

•250000 GPS probes/sec

Energy: Cooling

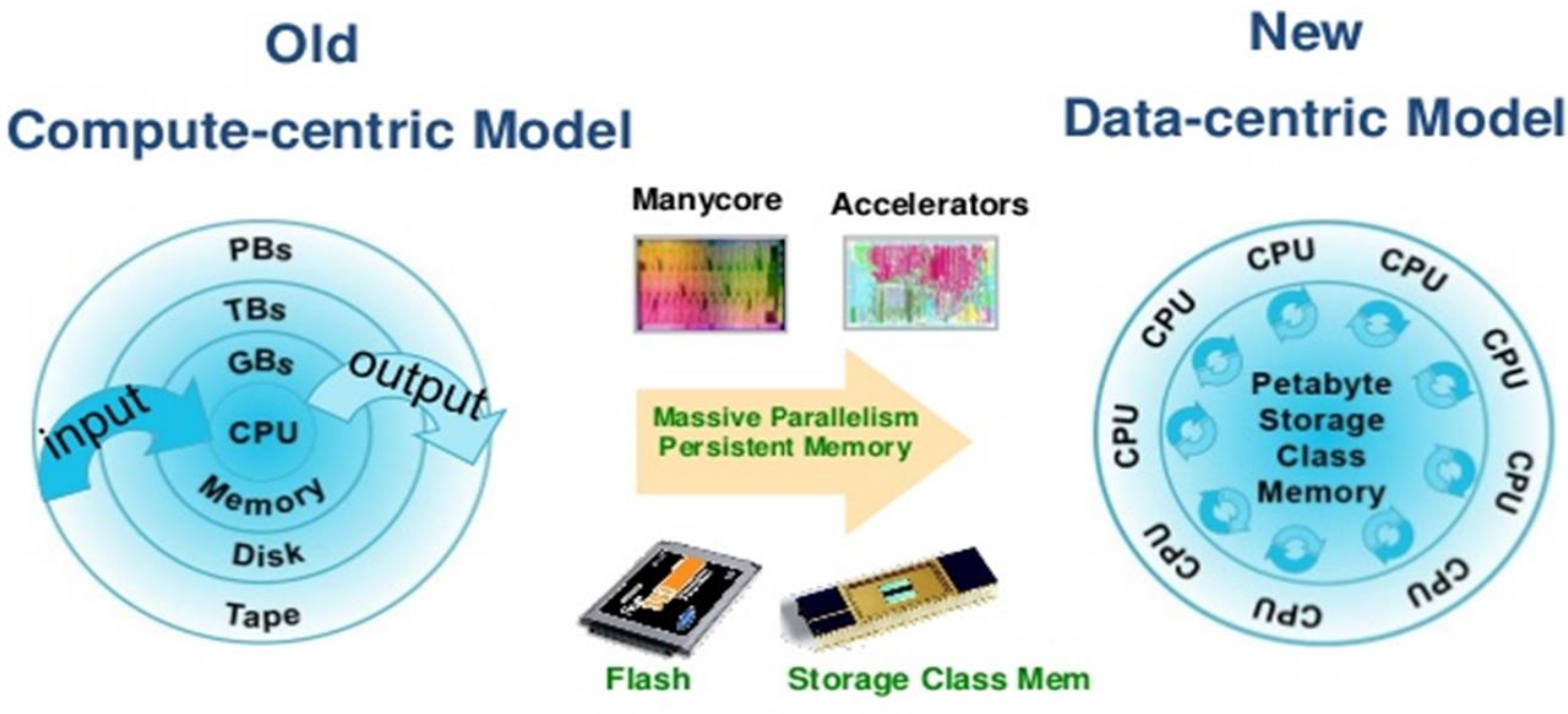
Energy: Processing and Clocks



- New VLSI Paradigms
- Asynchronous Hardware
- GALS: Globally Asynchronous Locally Synchronous

Source: Zhang and Theodoropoulos, SAMIPS, A Synthesisable MIPS Processor

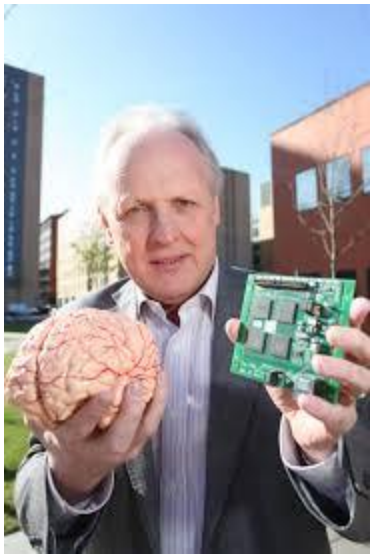
Trends in Computing



Source: IBM

SpiNNaker: ‘Human brain' supercomputer

- 57,600 **ARM9** processors
- each with 18 cores
- Total: 1,036,800 cores and over 7 TB of RAM



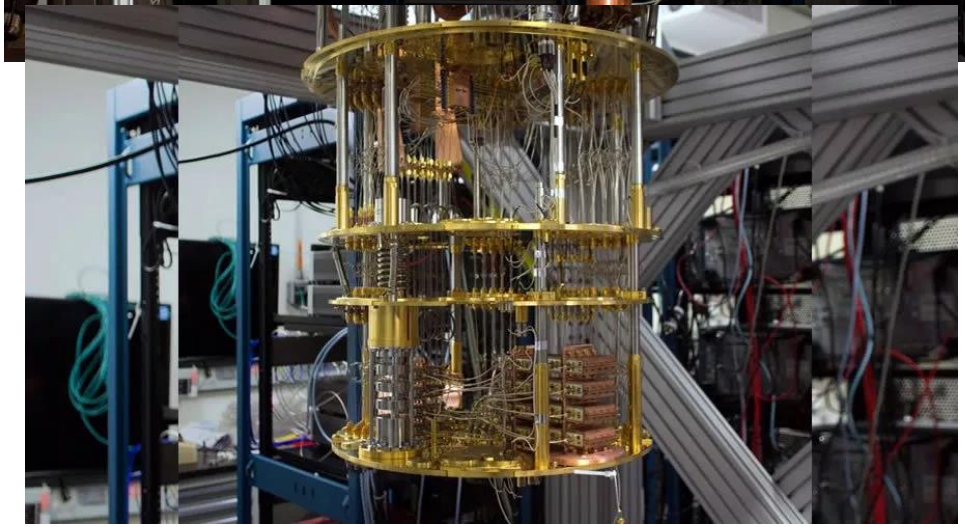
“SpiNNaker completely re-thinks the way conventional computers work. We’ve essentially created a machine that works more like a brain than a traditional computer”

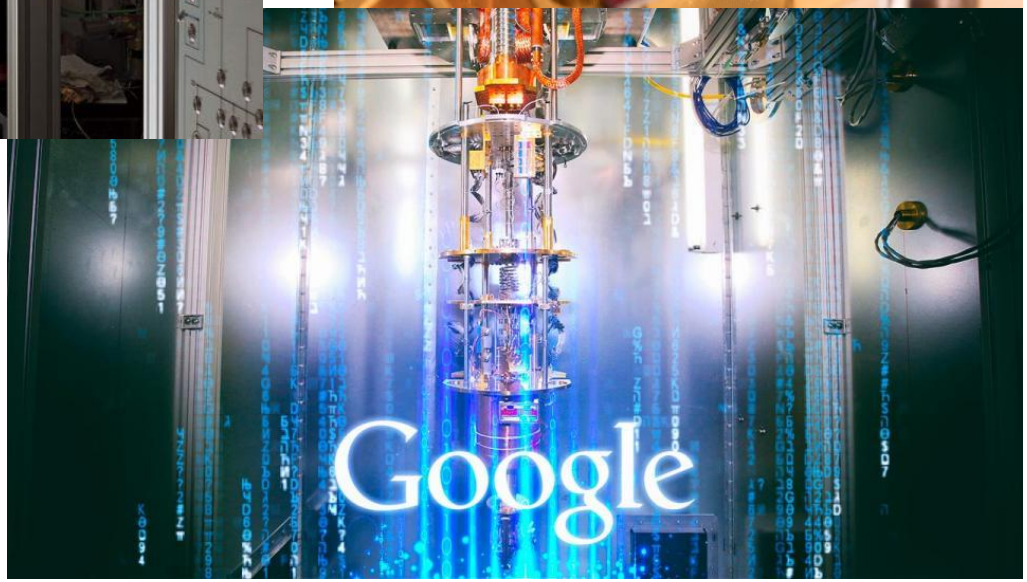
Steve Furber, ICL Professor of Computer Engineering, Manchester University


IBM Q

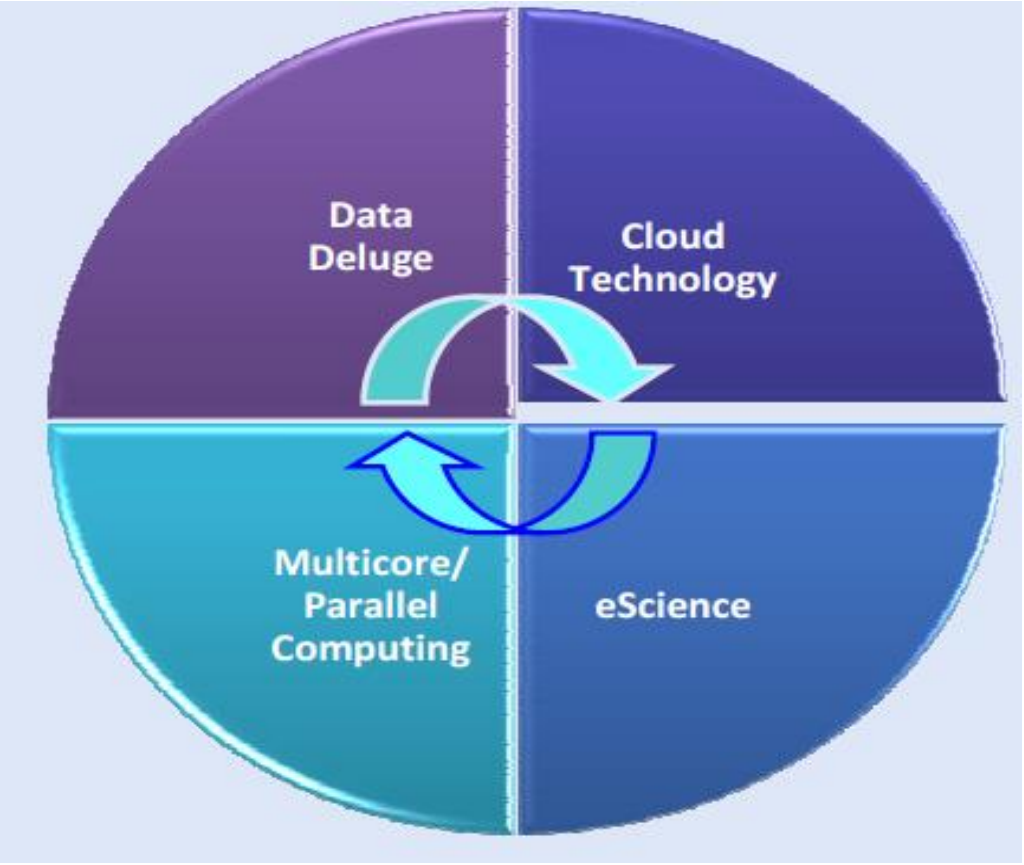

D-Wave 2000Q
D-Wave
D-Wave 2000Q


Google
Bristlecone




Google

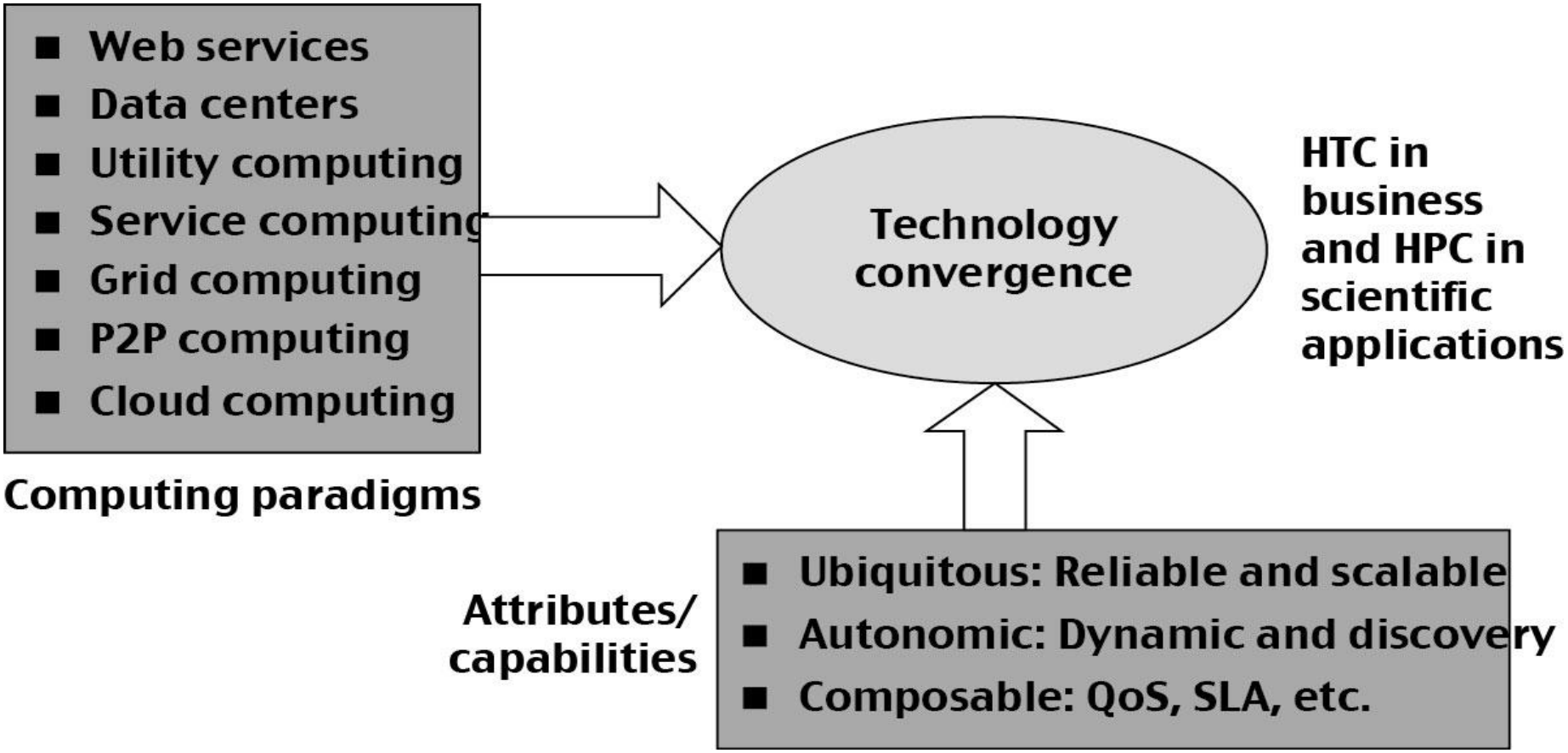
Interactions among 4 technical challenges : Data Deluge, Cloud Technology, eScience, and Multicore/Parallel Computing



(Courtesy of Judy Qiu, Indiana University, 2011)

From Desktop/HPC/Grids to Internet Clouds in 30 Years

- HPC moving from centralized supercomputers to geographically distributed desktops, desksides, clusters, and grids to clouds over last 30 years
- R/D efforts on HPC, clusters, Grids, P2P, and virtual machines has laid the foundation of cloud computing that has been greatly advocated since 2007
- Location of computing infrastructure in areas with lower costs in hardware, software, datasets, space, and power requirements – moving from desktop computing to datacenter-based clouds


Web services
Data centers
Utility computing
Service computing
Grid computing
P2P computing
Cloud computing
Computing paradigms
Technology convergence
HTC in business and HPC in scientific applications
Attributes/ capabilities
Ubiquitous: Reliable and scalable
Autonomic: Dynamic and discovery
Composable: QoS, SLA, etc.

A Typical Cluster Architecture

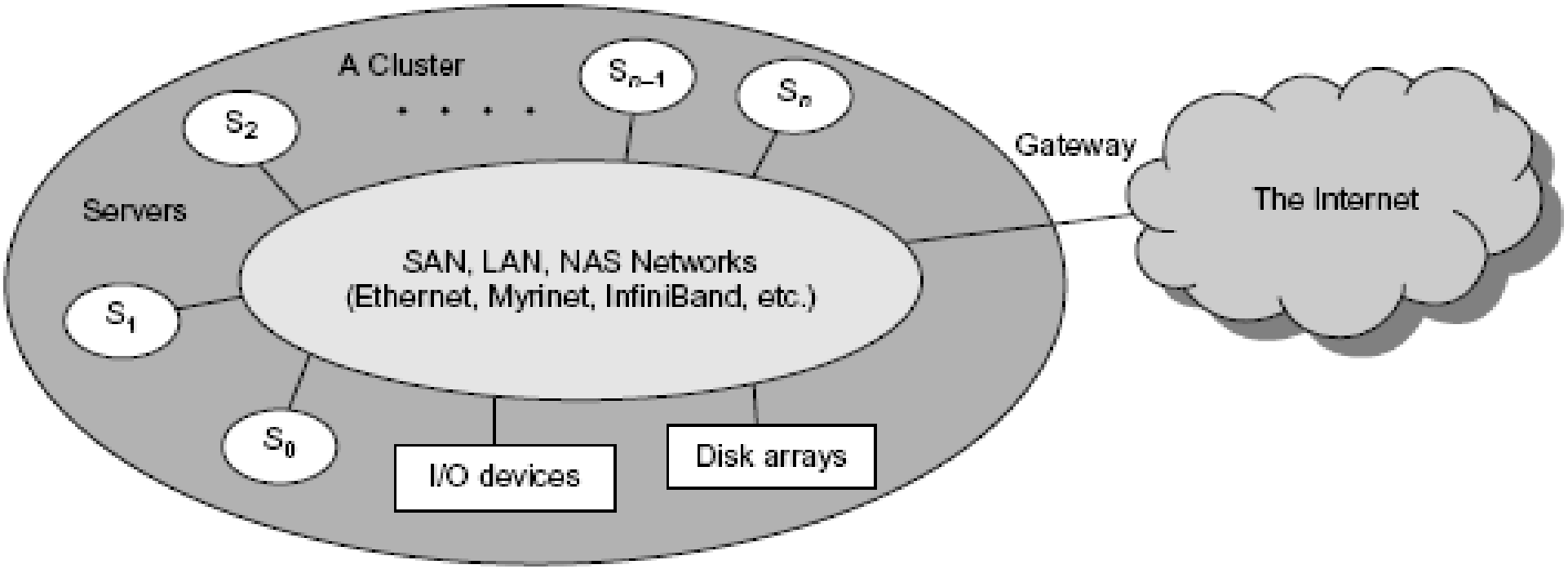


FIGURE 1.15

A cluster of servers interconnected by a high-bandwidth SAN or LAN with shared I/O devices and disk arrays; the cluster acts as a single computer attached to the Internet.

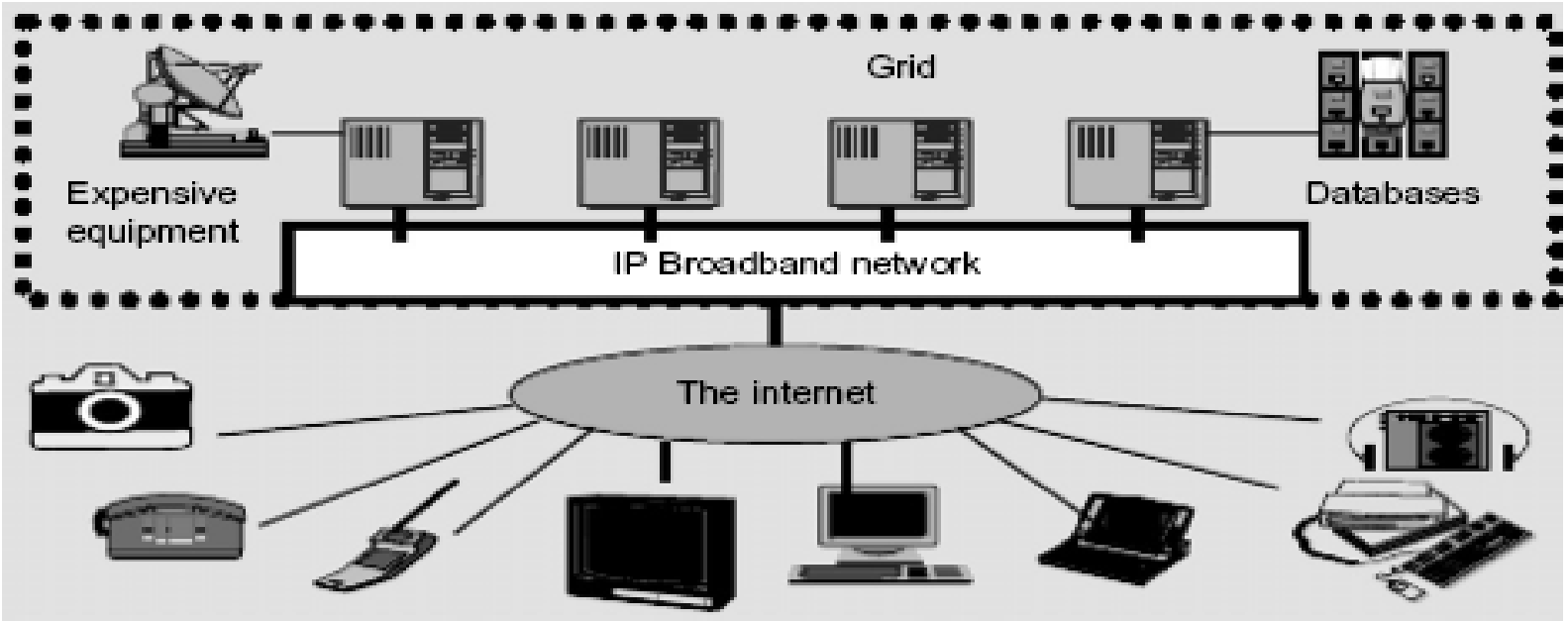


FIGURE 1.16

Computational grid or data grid providing computing utility, data and information services through resource sharing and cooperation among participating organizations.

A Typical Computational Grid

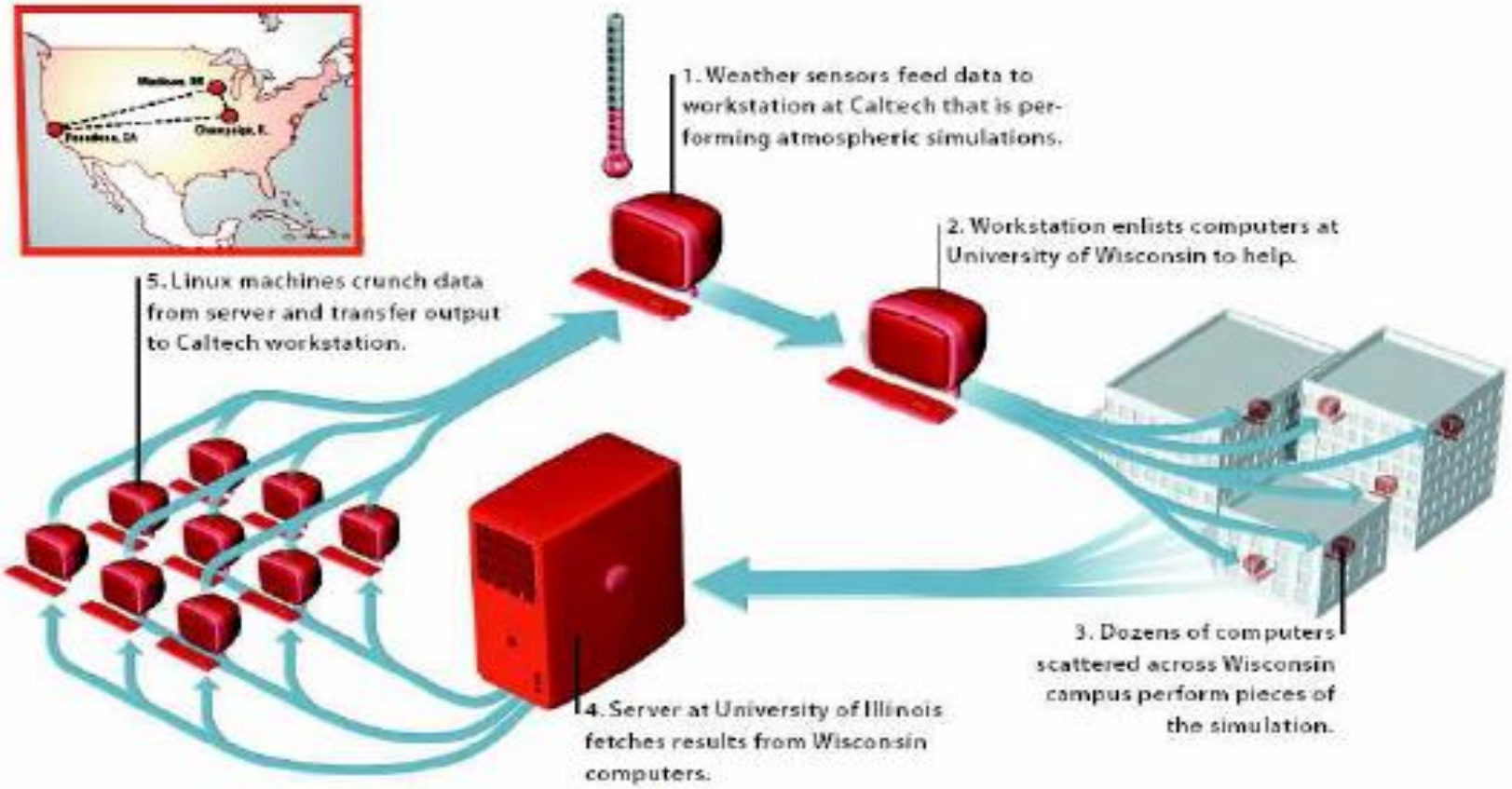


Figure 1.17 An example computational Grid built over specialized computers at three resource sites at Wisconsin, Caltech, and Illinois. (Courtesy of Michel Waldrop, "Grid Computing", IEEE *Computer Magazine*, 2000. [42])

Grid Standards and Middleware :

Table 1.9 Grid Standards and Toolkits for scientific and Engineering Applications

Grid Standards	Major Grid Service Functionalities	Key Features and Security Infrastructure
OGSA Standard	Open Grid Service Architecture offers common grid service standards for general public use	Support heterogeneous distributed environment, bridging CA, multiple trusted intermediaries, dynamic policies, multiple security mechanisms, etc.
Globus Toolkits	Resource allocation, Globus security infrastructure (GSI), and generic security service API	Sign-in multi-site authentication with PKI, Kerberos, SSL, Proxy, delegation, and GSS API for message integrity and confidentiality
IBM Grid Toolbox	AIX and Linux grids built on top of Globus Toolkit, autonomic computing, Replica services	Using simple CA, granting access, grid service (ReGS), supporting Grid application for Java (GAF4J), GridMap in IntraGrid for security update.

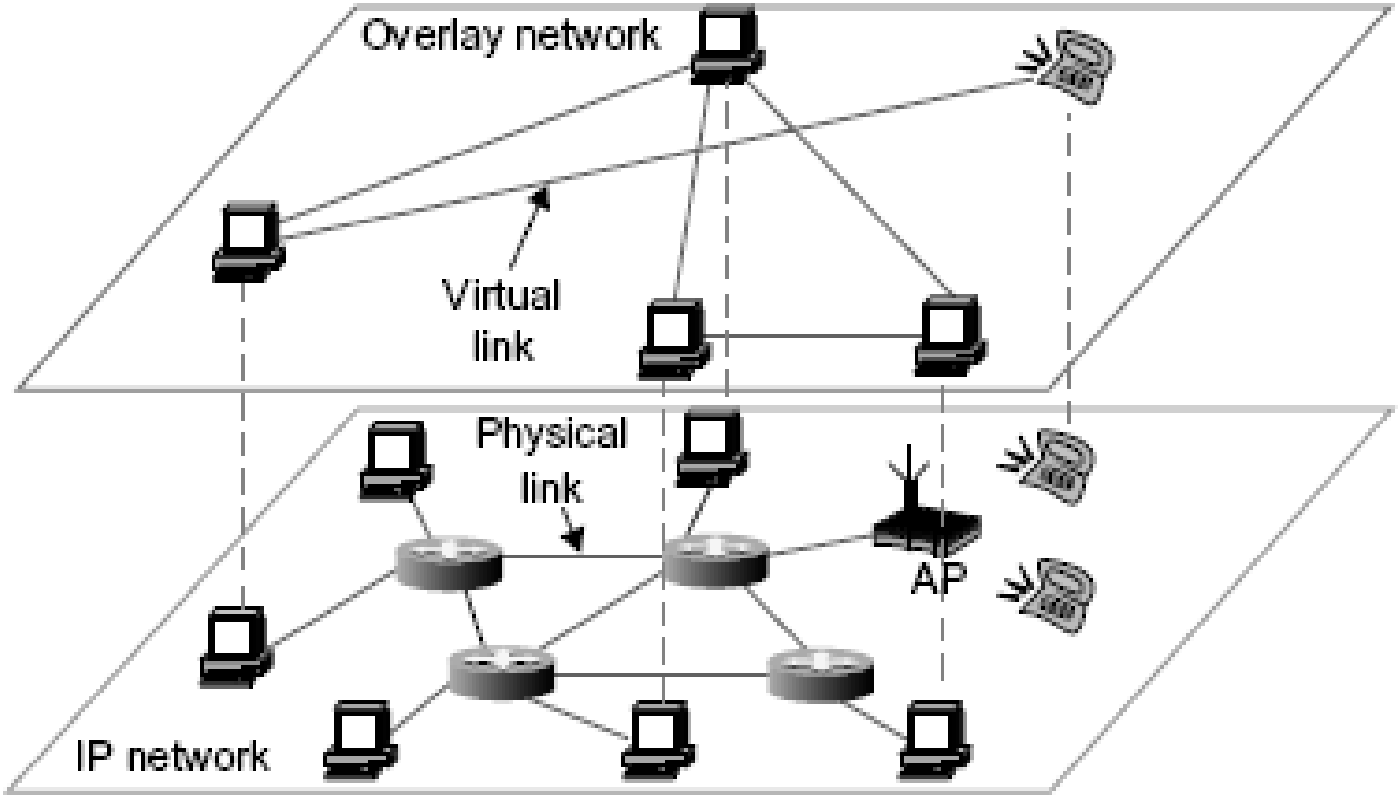


FIGURE 1.17

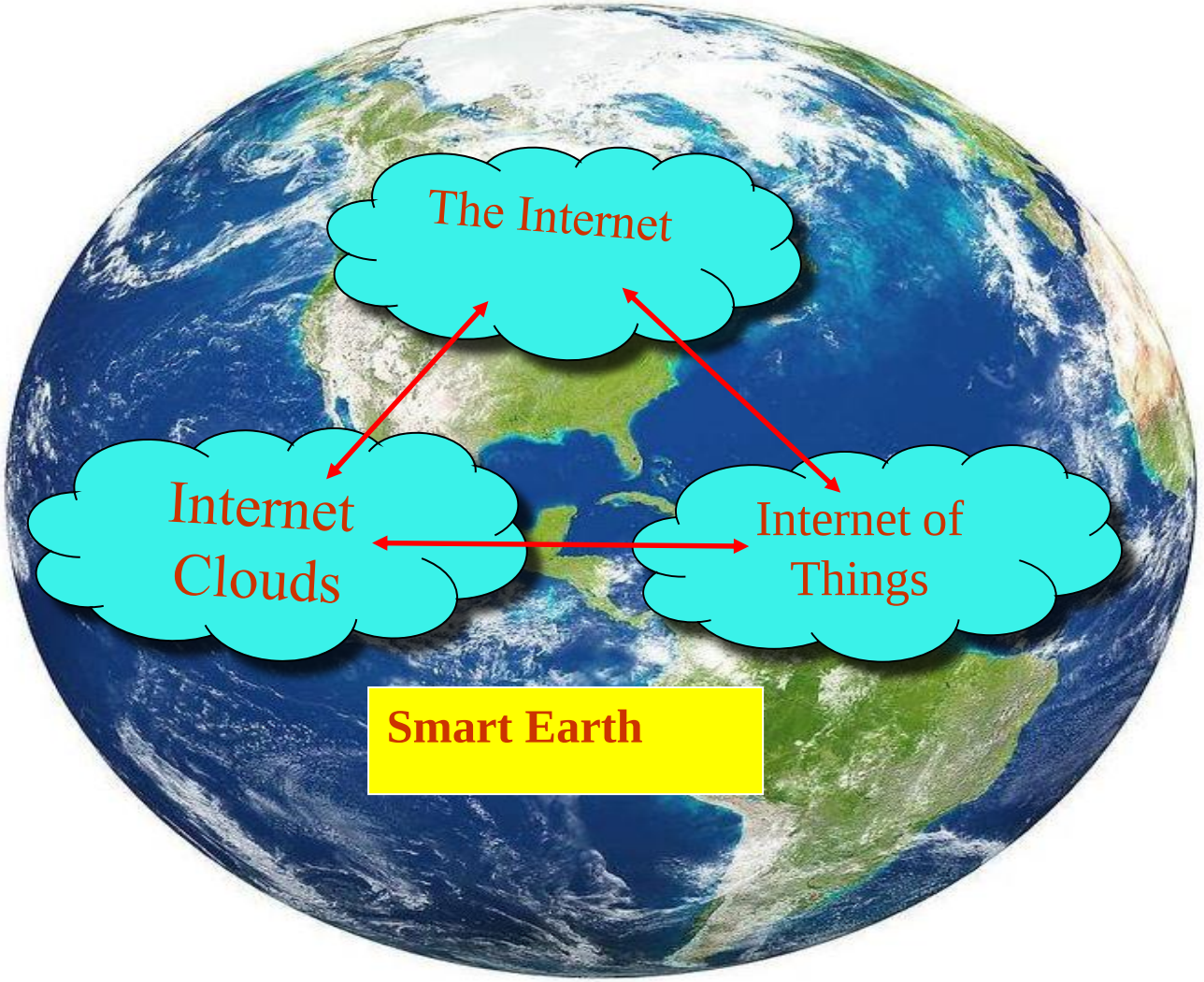
The structure of a P2P System by mapping a physical IP network to an overlay network built with virtual Links.

(Courtesy of Zhenyu Li, Institute of Computing Technology, Chinese Academy of Sciences, 2

Table 1.5 Major Categories of P2P Network Families [42]

System Features	Distributed File Sharing	Collaborative Platform	Distributed P2P Computing	P2P Platform
Attractive Applications	Content distribution of MP3 music, video, open software, etc.	Instant messaging, collaborative design and gaming	Scientific exploration and social networking	Open networks for public resources
Operational Problems	Loose security and serious online copyright violations	Lack of trust, disturbed by spam, privacy, and peer collusion	Security holes, selfish partners, and peer collusion	Lack of standards or protection protocols
Example Systems	Gnutella, Napster, eMule, BitTorrent, Aimster, KaZaA, etc.	ICQ, AIM, Groove, Magi, Multiplayer Games, Skype, etc.	SETI@home, Geonome@home, etc.	JXTA, .NET, FightingAid@home, etc.

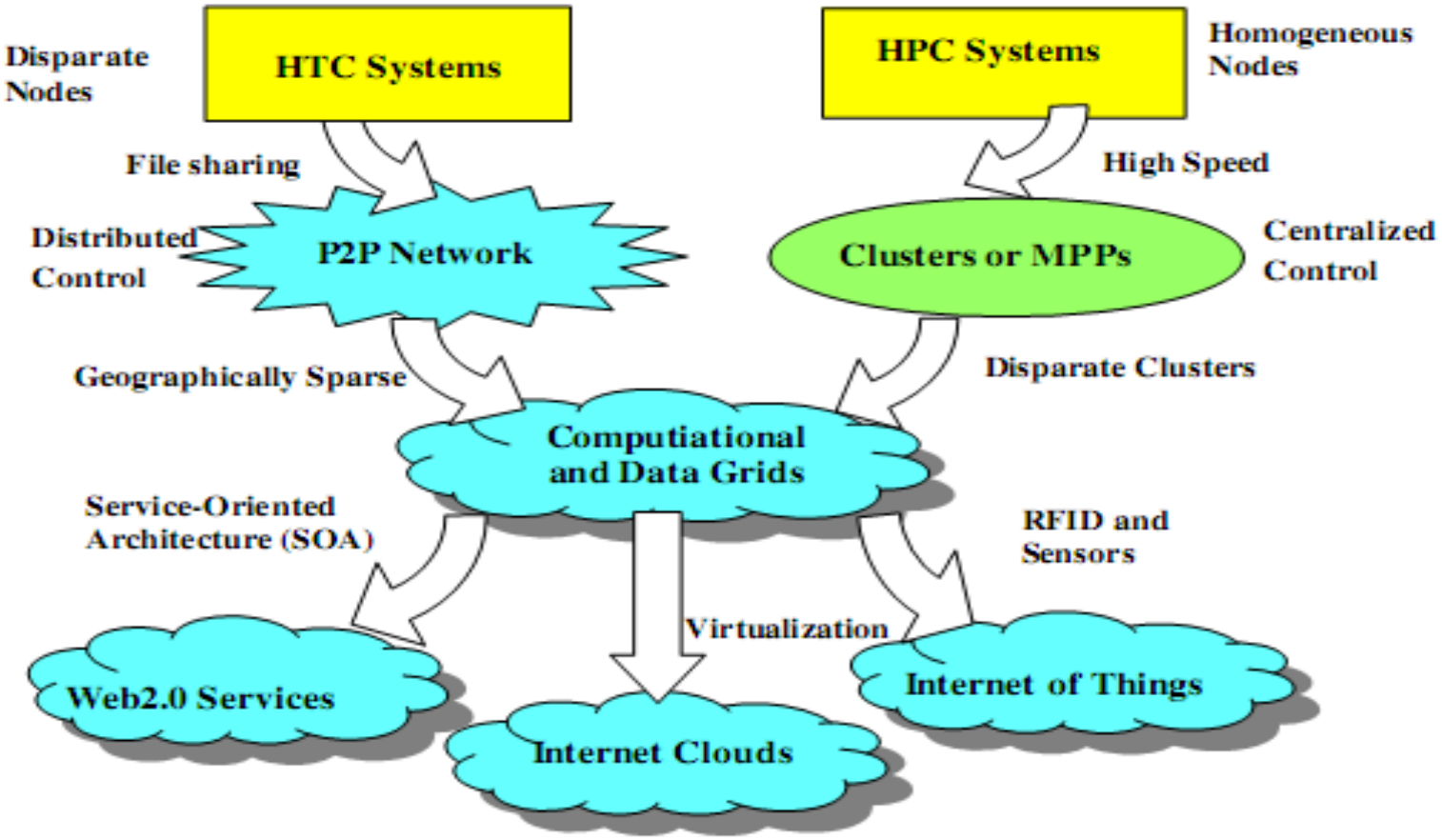
The Internet of Things (IoT)



Smart Earth:

An IBM Dream

Clouds and Internet of Things



HPC: High-Performance Computing

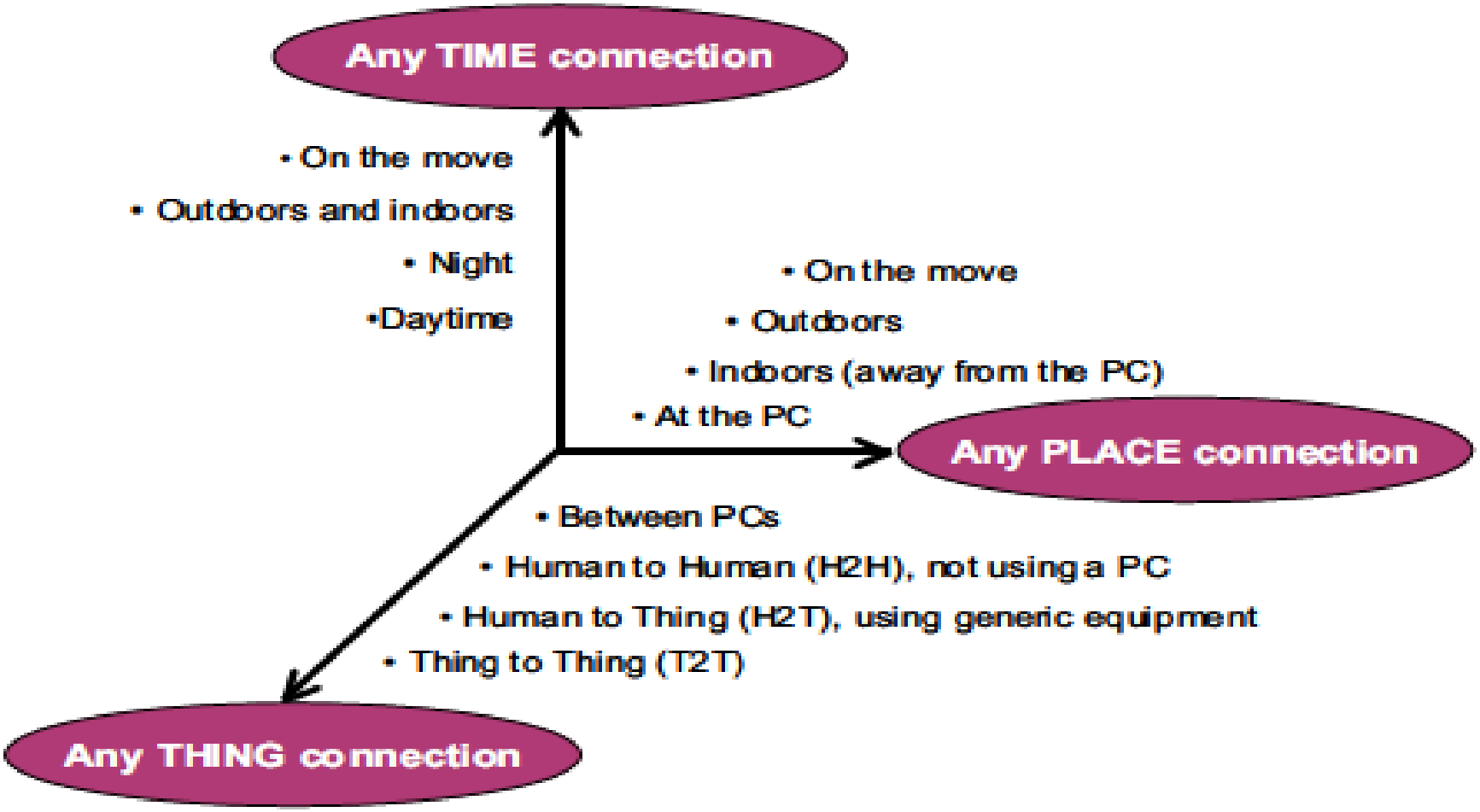
HTC: High-Throughput Computing

P2P: Peer to Peer

MPP: Massively Parallel Processors

Source: **K. Hwang, G. Fox, and J. Dongarra,** ***Distributed and Cloud Computing,*** **Morgan Kaufmann, 2012.**

Opportunities of IoT in 3 Dimensions



(courtesy of Wikipedia, 2010)

Table 1.2 Classification of Distributed Parallel Computing Systems

Functionality, Applications	Multicomputer Clusters [27, 33]	Peer-to-Peer Networks [40]	Data/Computational Grids [6, 42]	Cloud Platforms [1, 9, 12, 17, 29]
Architecture, Network Connectivity and Size	Network of compute nodes interconnected by SAN, LAN, or WAN, hierarchically	Flexible network of client machines logically connected by an overlay network	Heterogeneous clusters interconnected by high-speed network links over selected resource sites.	Virtualized cluster of servers over datacenters via service-level agreement
Control and Resources Management	Homogeneous nodes with distributed control, running Unix or Linux	Autonomous client nodes, free in and out, with distributed self-organization	Centralized control, server oriented with authenticated security, and static resources	Dynamic resource provisioning of servers, storage, and networks over massive datasets
Applications and network-centric services	High-performance computing, search engines, and web services, etc.	Most appealing to business file sharing, content delivery, and social networking	Distributed super-computing, global problem solving, and datacenter services	Upgraded web search, utility computing, and outsourced computing services
Representative Operational Systems	Google search engine, SunBlade, IBM Road Runner, Cray XT4, etc.	Gnutella, eMule, BitTorrent, Napster, KaZaA, Skype, JXTA, and .NET	TeraGrid, GriPhyN, UK EGEE, D-Grid, ChinaGrid, etc.	Google App Engine, IBM Bluecloud, Amazon Web Service(AWS), and Microsoft Azure,

Parallel and Distributed Programming

Table 1.7 Parallel and Distributed Programming Models and Tool Sets

Model	Description	Features
MPI	A library of subprograms that can be called from C or FORTRAN to write parallel programs running on distributed computer systems [6,28,42]	Specify synchronous or asynchronous point-to-point and collective communication commands and I/O operations in user programs for message-passing execution
MapReduce	A Web programming model for scalable data processing on large clusters over large data sets, or in Web search operations [16]	*Map* function generates a set of intermediate key/value pairs; *Reduce* function merges all intermediate values with the same key
Hadoop	A software library to write and run large user applications on vast data sets in business applications (http://hadoop.apache.org/core)	A scalable, economical, efficient, and reliable tool for providing users with easy access of commercial clusters

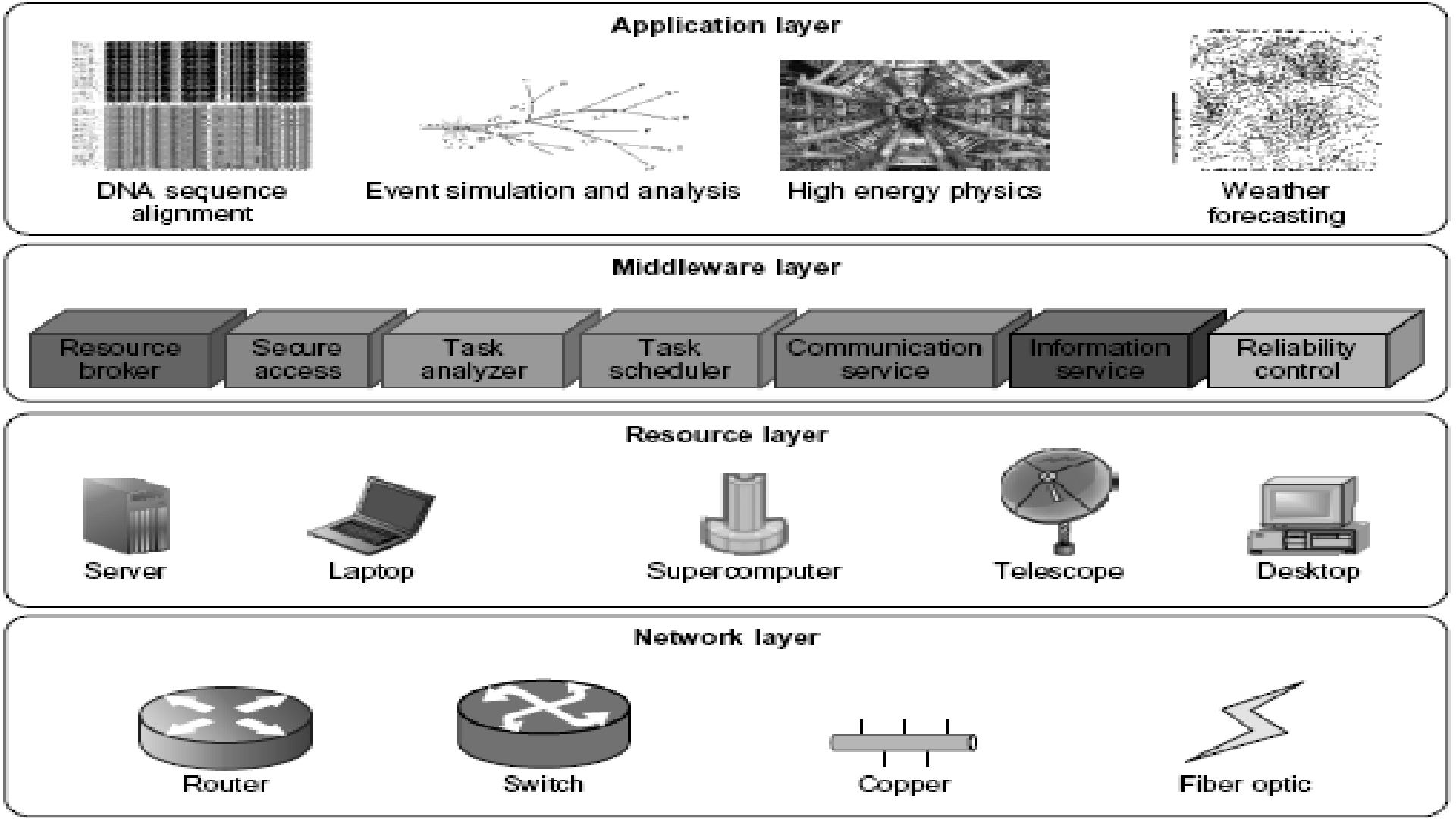


FIGURE 1.26

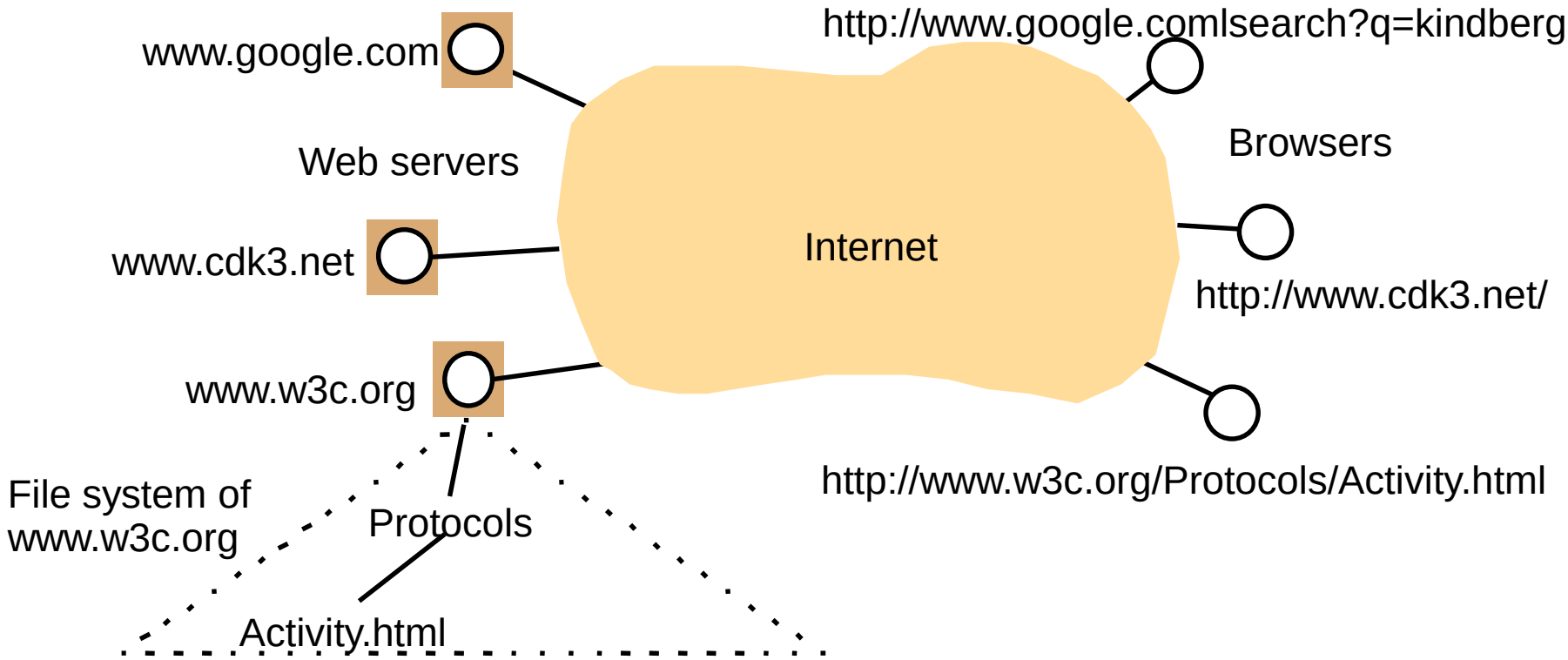
Four operational layers of distributed computing systems.

(Courtesy of Zomaya, Rivandi and Lee of the University of Sydney [33])

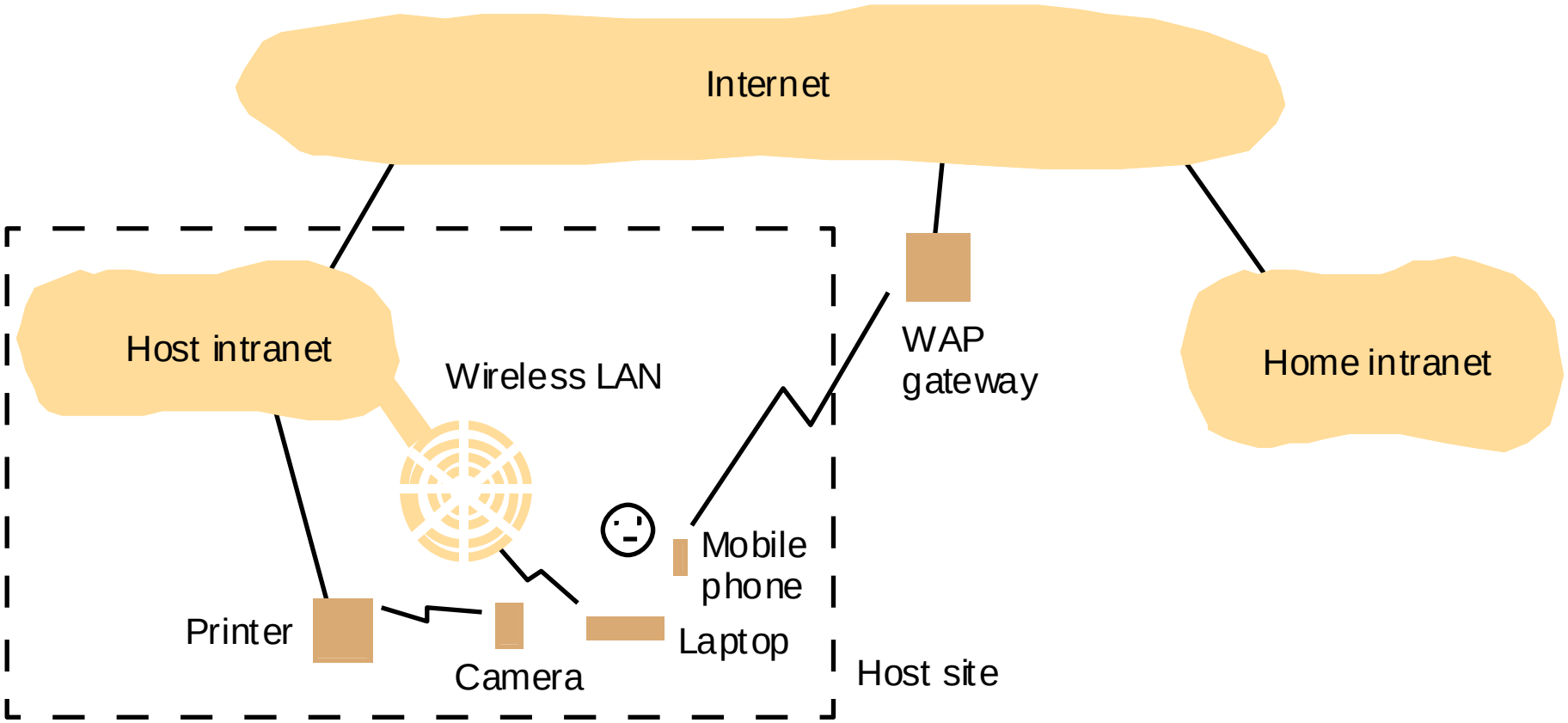
Distributed Systems: Definitions

- What is a distributed system: a system in which components located in networked computers communicate and coordinate their actions only by passing messages.
- Motivation: sharing of resources

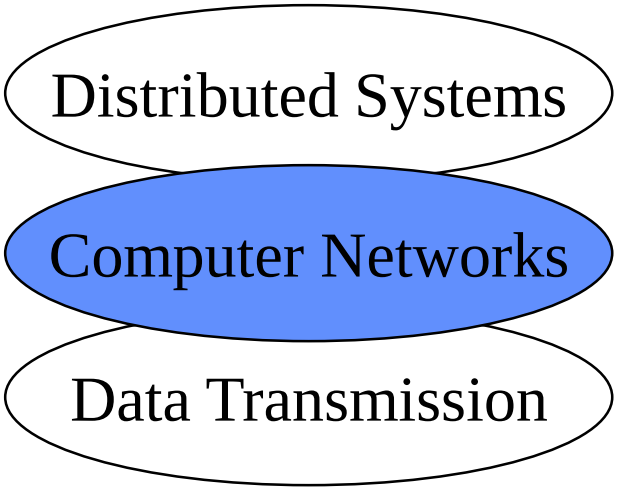
Resource sharing: Web servers and web browsers



Portable and handheld devices in a distributed system

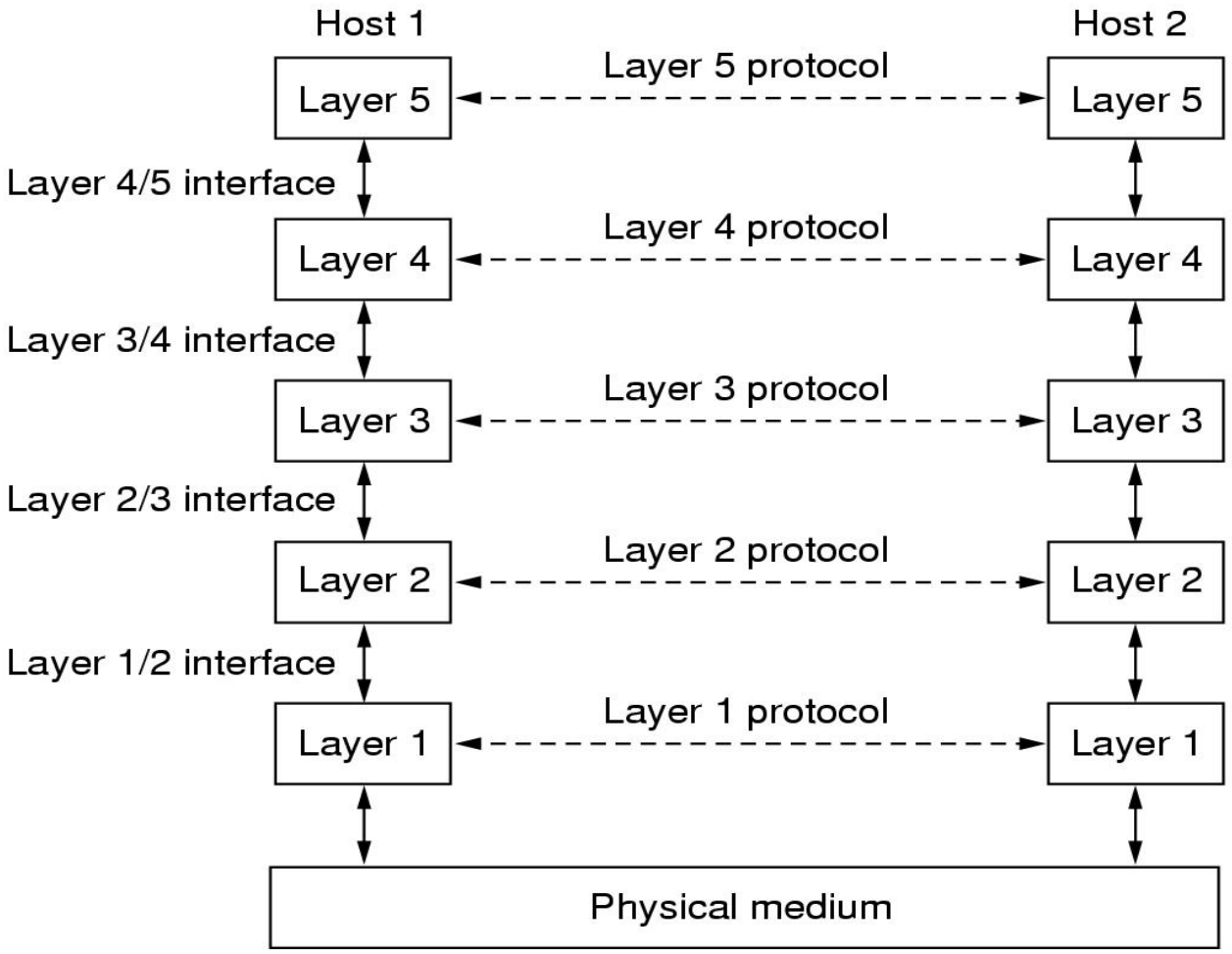


- Computer Network: A collection of *interconnected autonomous* computers
  - Generality: Built from general purpose hardware - not optimised for any particular application or data type
- Computer Network vs. Distributed System
  - Transparency
  - DS is a software system that runs on top of CN



CN Layered Architecture

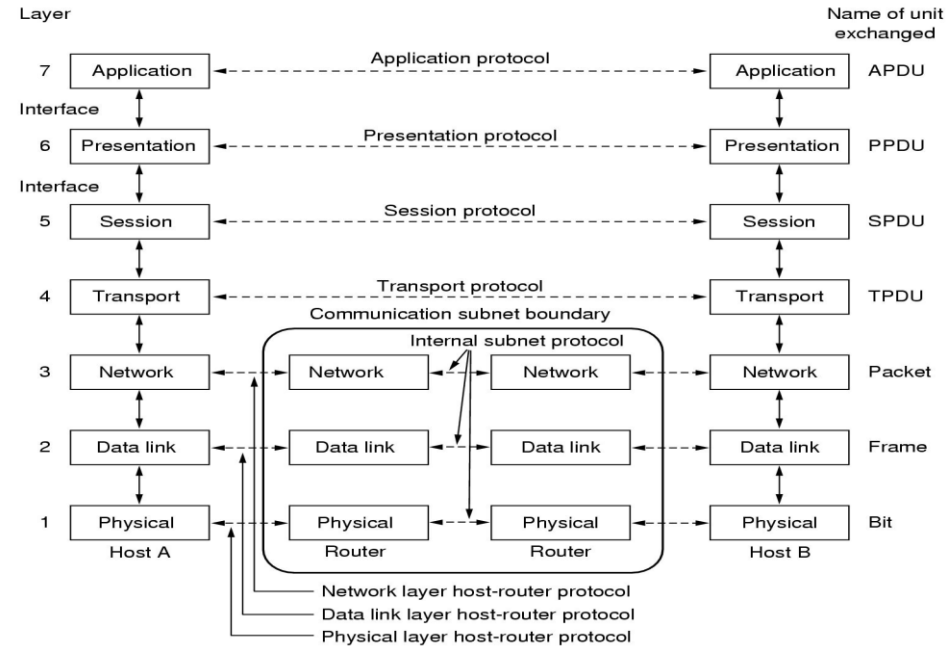
- The purpose of each layer is to offer a communication services to higher level layers
- Each layer has two interfaces
  - peer-to-peer interface
    - defines the form and types of messages exchanged between peers (indirect communication)
  - service interface
    - defines the primitives (operations) that a layer provides to the layer above it
- Layering is non-linear

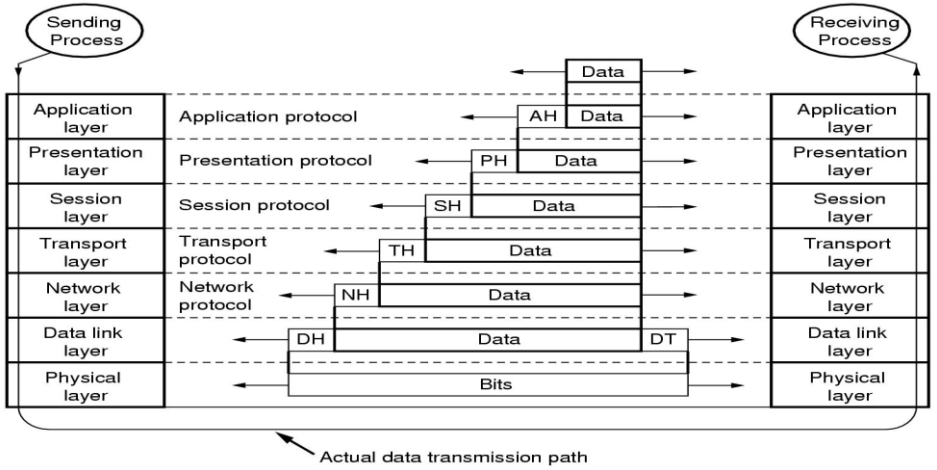


ISO OSI Architecture

- International Standards Organisation (ISO)
- *Physical*: transmission of raw bits onto the communications medium
- *Data link*: reliable transmission of frames, flow control, arbitration
- *Network*: packet switching, routing congestion control
- *Transport*: process-to-process channel, node-to-node connection, provides user services, flow control, multiplexing
- *Session*
- *Presentation*
- *Application*

ISO OSI Architecture





OSI Session Layer

- The protocols necessary to establish and maintain a connection or session between 2 end-users
- Transport vs session

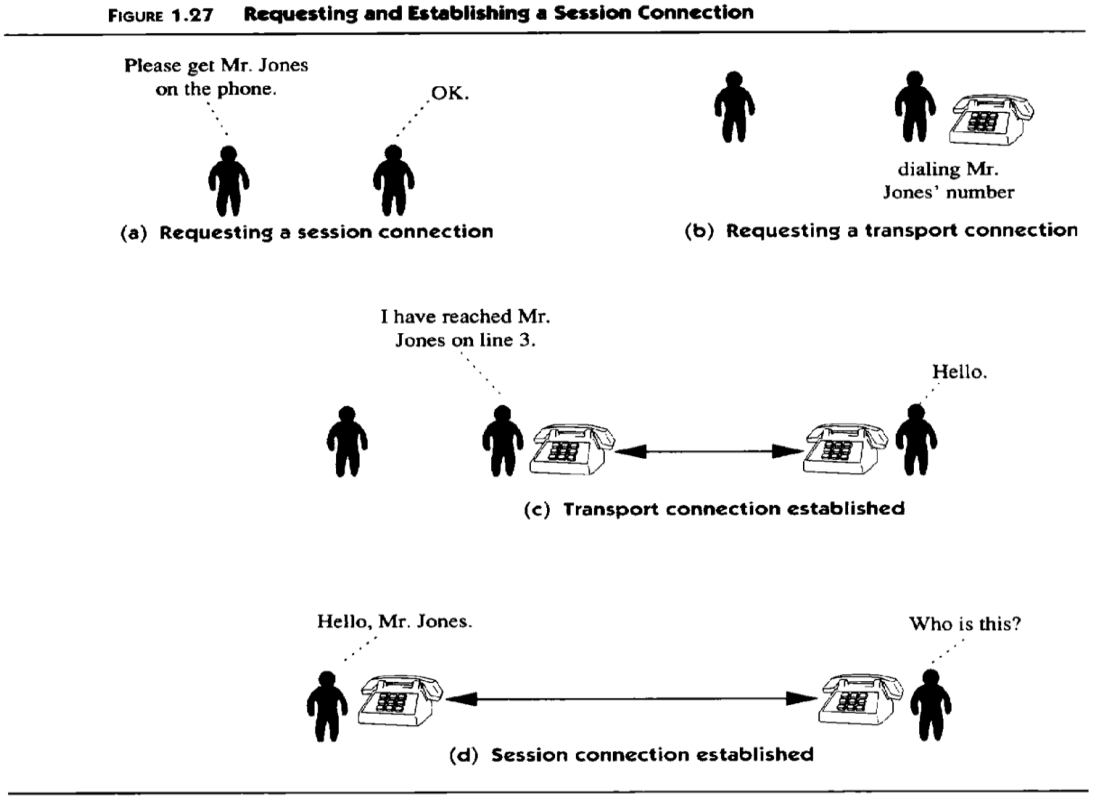
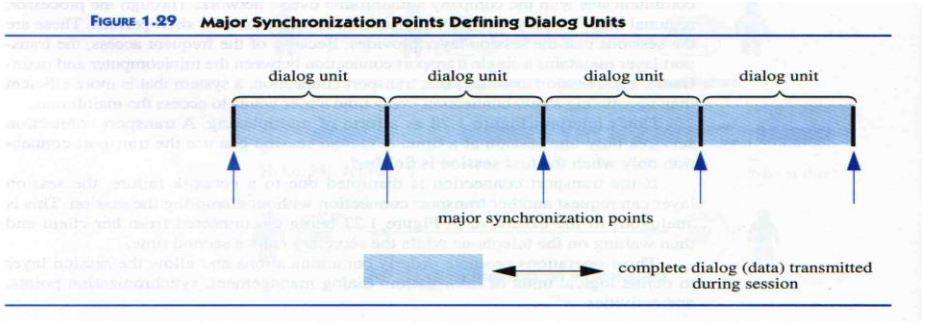
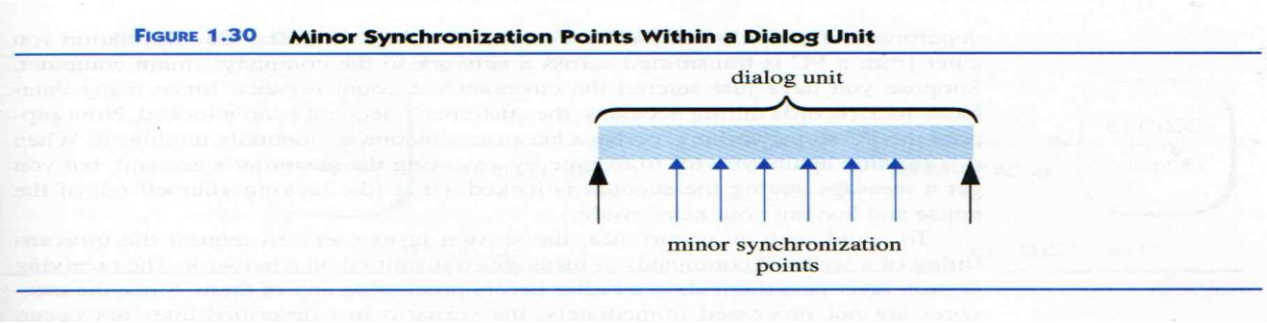


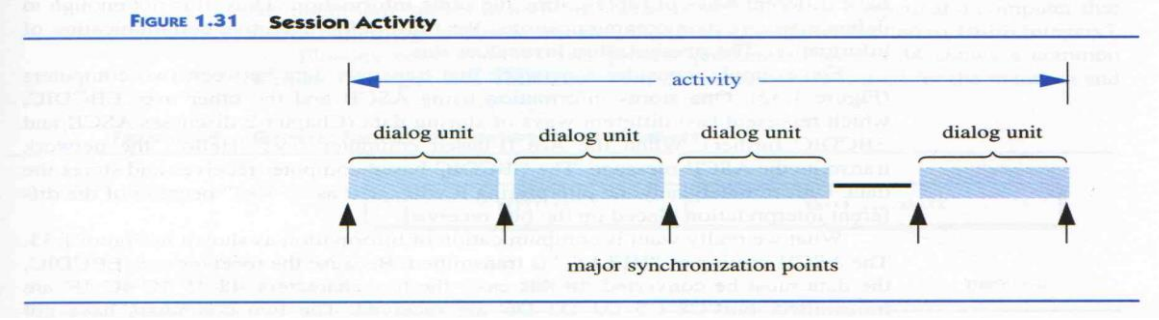
FIGURE 1.27 Requesting and Establishing a Session Connection

OSI Session Layer

- Orderly communication
  - dialog management
  - synchronisation points (major and minor)
  - activities



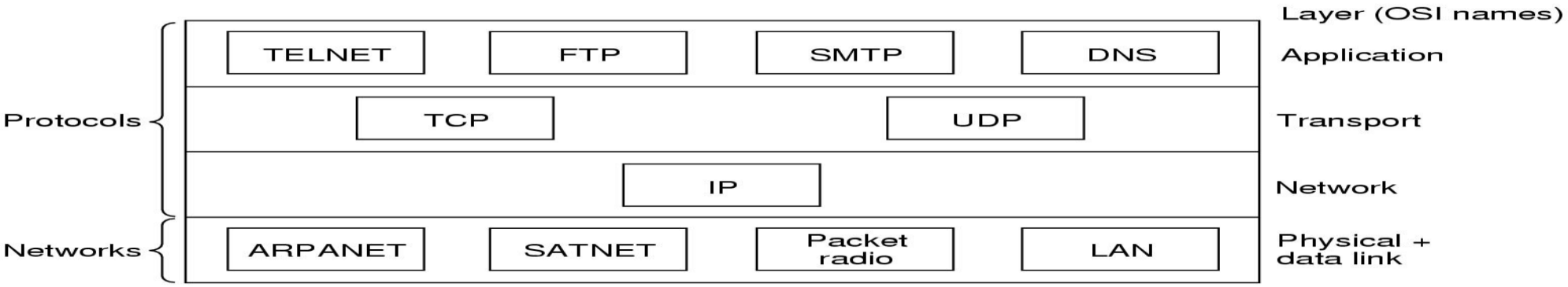




OSI Presentation & Application Layers

- Presentation
  - Effective communication of information rather than of data
  - Code and number conversion
  - Transmission of sophisticated data structures
  - Data compression
- Application
  - Electronic mail
  - File transfer protocols (ftp)
  - virtual terminal protocols (telnet)
  - Distributed system (distributed database)
  - Client-server

Internet Architecture (TCP/IP)



- Host-to-Network Layer (OSI Physical and Data link layers)
- Internet Layer (OSI Network layer - Internet Protocol/IP)
- Transport Layer (Transmission Control Protocol/TCP & User Datagram Protocol/UDP)
- Application Layer

Challenges

- Heterogeneity (mobile code)
- Openness
- Security (denial of service, security of mobile code etc.)
- Scalability (cost of physical resources, performance loss, availability, bottlenecks)
- Failure (detection, correct/hide, tolerate, recover, redundancy)
- Concurrency (consistency)
- Transparency

Transparencies

Access transparency: enables local and remote resources to be accessed using identical operations.
Location transparency: enables resources to be accessed without knowledge of their physical or network location (for example, which building or IP address).
Concurrency transparency: enables several processes to operate concurrently using shared resources without interference between them.
Replication transparency: enables multiple instances of resources to be used to increase reliability and performance without knowledge of the replicas by users or application programmers.
Failure transparency: enables the concealment of faults, allowing users and application programs to complete their tasks despite the failure of hardware or software components.
Mobility transparency: allows the movement of resources and clients within a system without affecting the operation of users or programs.
Performance transparency: allows the system to be reconfigured to improve performance as loads vary.
Scaling transparency: allows the system and applications to expand in scale without change to the system structure or the application algorithms.