Windows NT Scalability

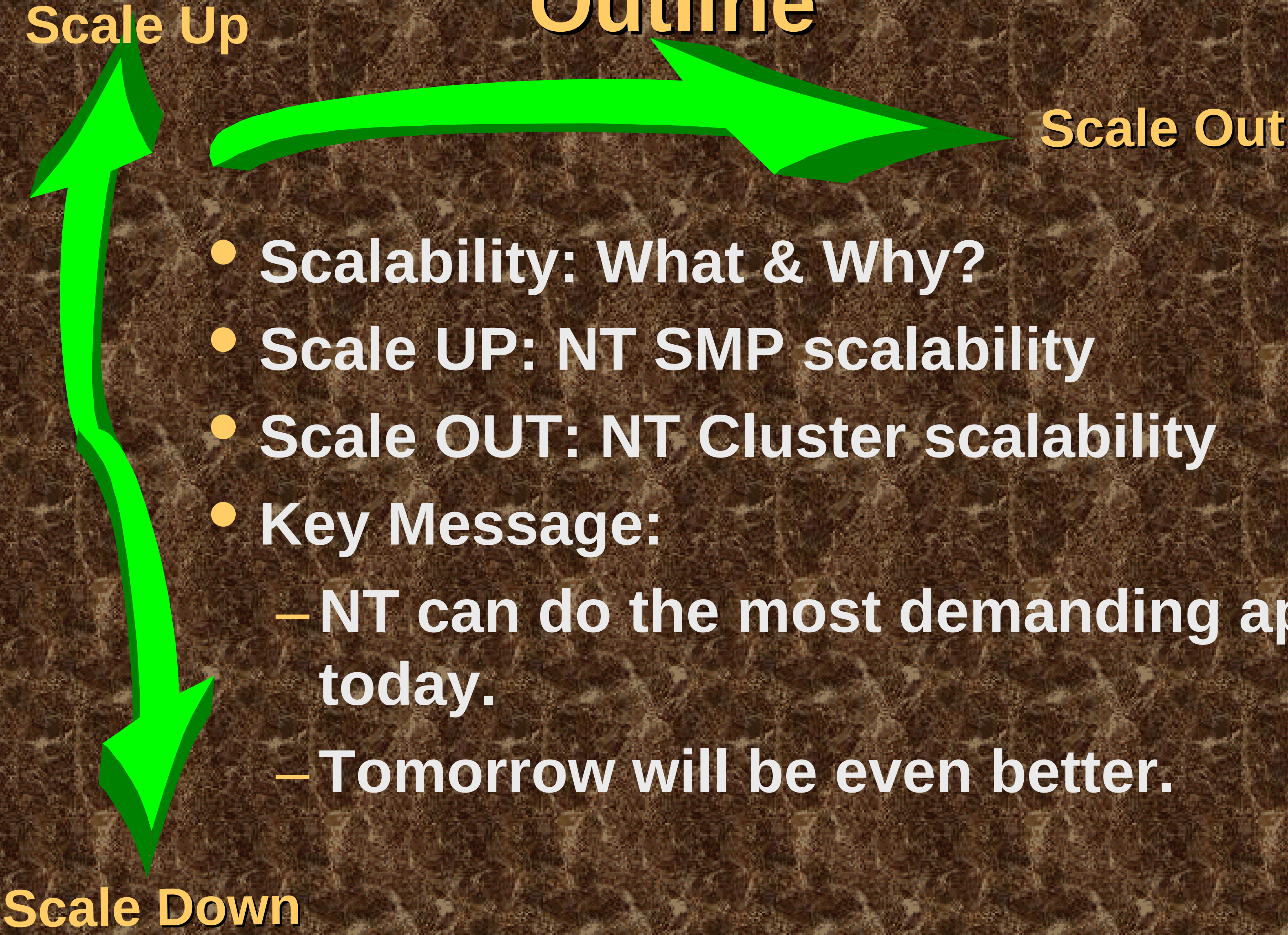
Jim Gray

Microsoft Research

Gray@Microsoft.com

http/www.research.Microsoft.com/~Gray/talks/

Outline



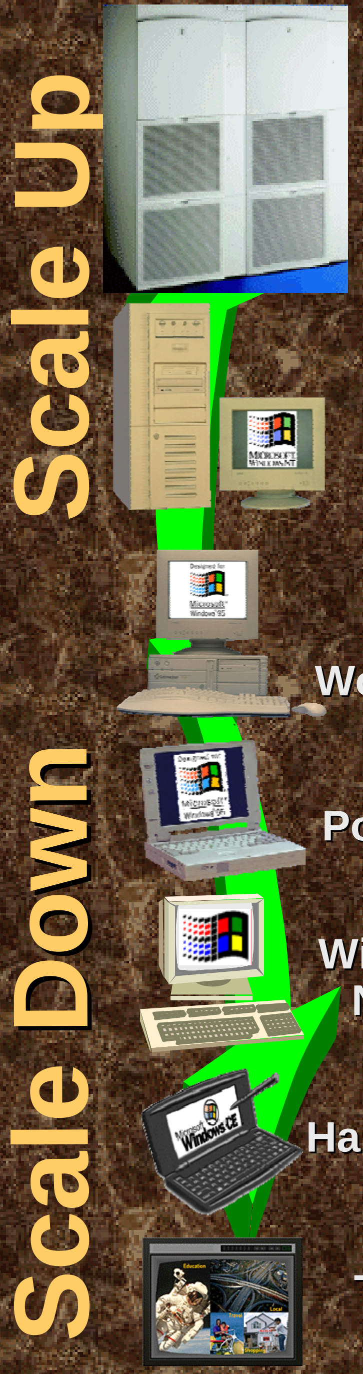
Scale Up

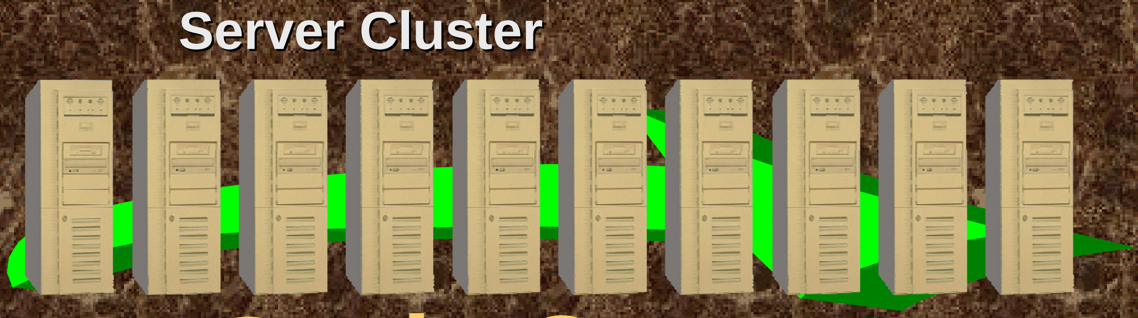
Scale Out

Scale Down

- Scalability: What & Why?
- Scale UP: NT SMP scalability
- Scale OUT: NT Cluster scalability
- Key Message:
  - NT can do the most demanding apps today.
  - Tomorrow will be even better.

What is Scalability?

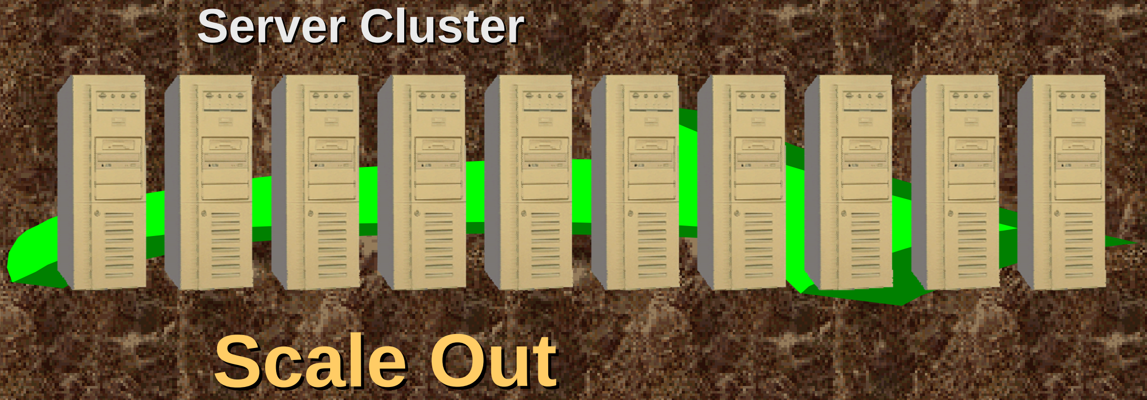


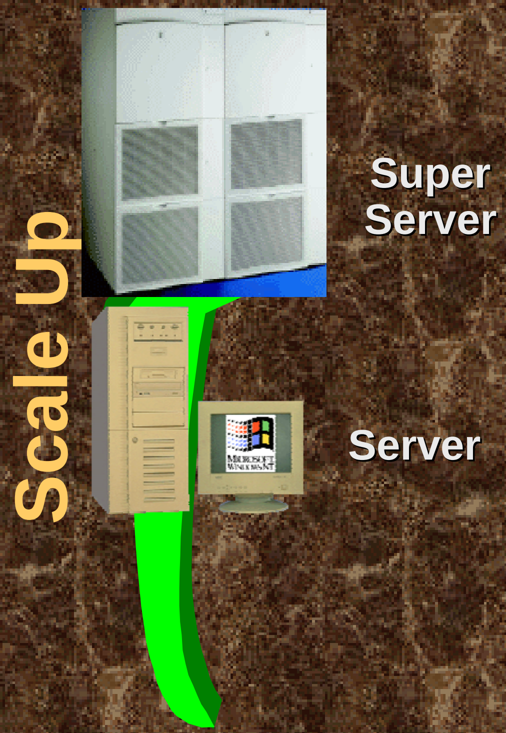


Scale Out

- Grow without limits
  – Capacity
  – Throughput
- Do not add complexity
  – design
  – administer
  – Operate
  – Use

Scale UP & OUT Focus Here





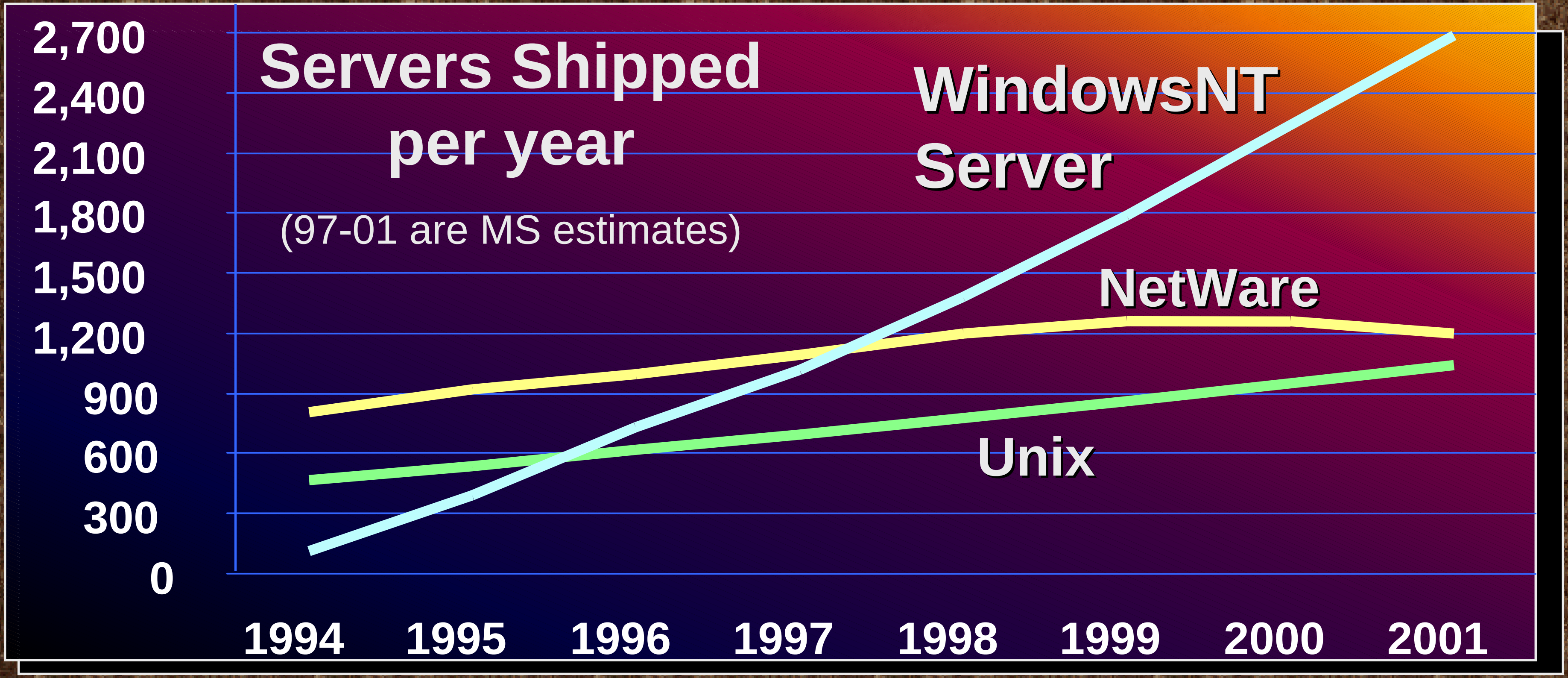
- **Grow without limits**
  - SMP: 4, 8, 16, 32 CPUs
  - 64-bit addressing
  - Huge storage
- **Cluster Requirements**
  - Auto manage
  - High availability
  - Transparency
  - Programming tools & **apps**

Scalability is Important

- Automation benefits growing
  - ROI of 1 month….
- Slice price going to zero
  - Cyberbrick costs 5k$
- Design, Implement & Manage cost going down
  - DCOM & Viper make it easy!
  - NT Clusters are easy!
- Billions of clients imply millions of HUGE servers.
- Thin clients imply huge servers.

Q: Why Does Microsoft Care?

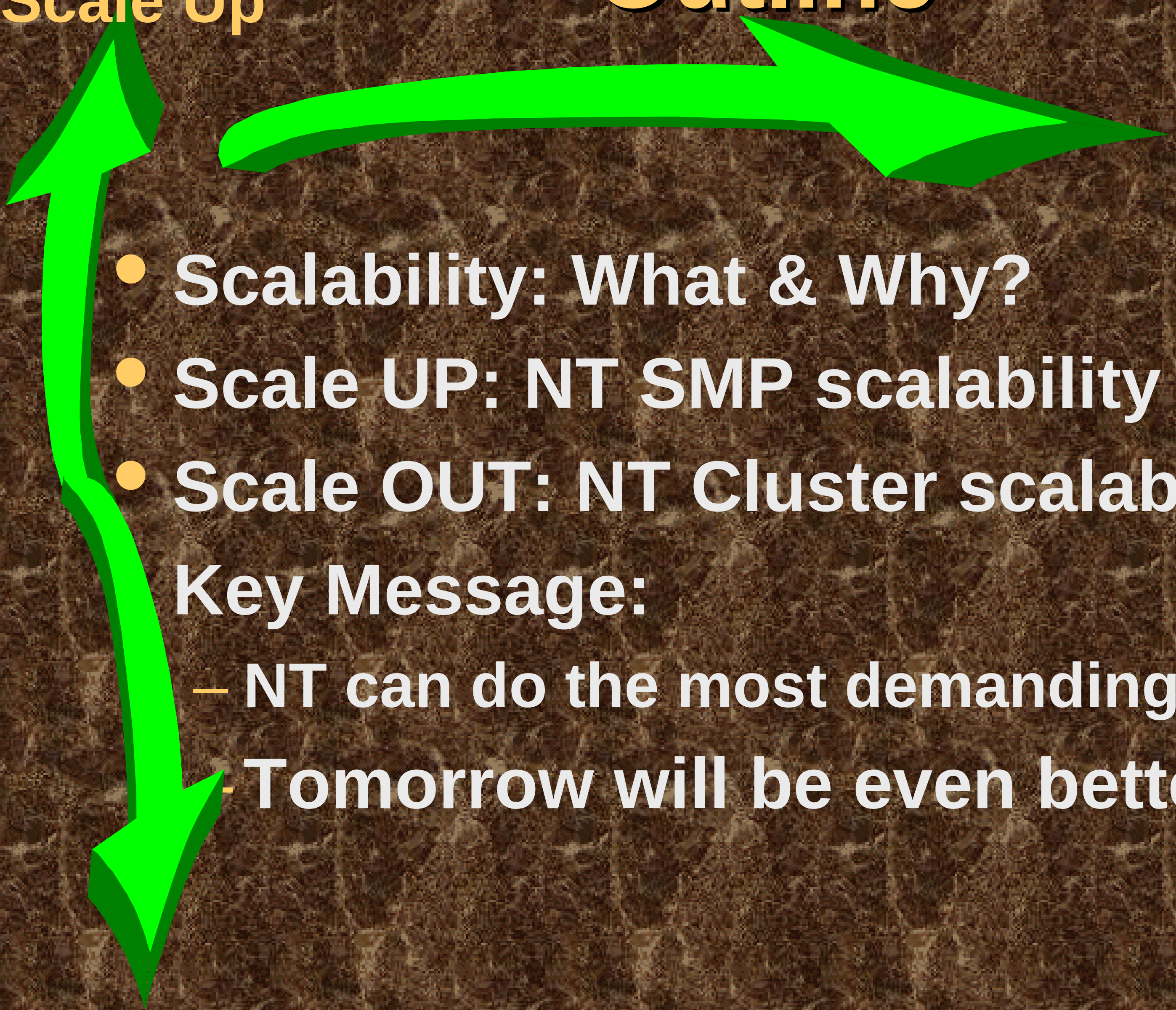
A: Billions of clients need millions of servers



Expect Microsoft to work hard on
Scaleable Windows NT and
Scaleable BackOffice.

Key technique: *INTEGRATION.*

Outline



Scale Up

Scale Out

Scale Down

- Scalability: What & Why?
- Scale UP: NT SMP scalability
- Scale OUT: NT Cluster scalability

Key Message:

- NT can do the most demanding apps today.
- Tomorrow will be even better.

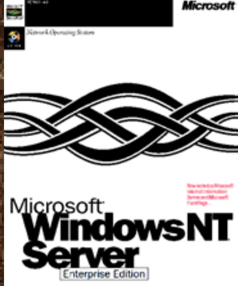
How Scaleable is NT??
The Single Node Story

- 64 bit file system in NT 1, 2, 3, 4, 5
- 8 node SMP in NT 4.E, 32 node OEM
- 64 bit addressing in NT 5
- 1 Terabyte SQL Databases (PetaByte capable)
- 10,000 users (TPC-C benchmark)
- 100 Million web hits per day (IIS)
- 50 GB Exchange mail store

  next release designed for 16 TB
- 50,000 POP3 users on Exchange

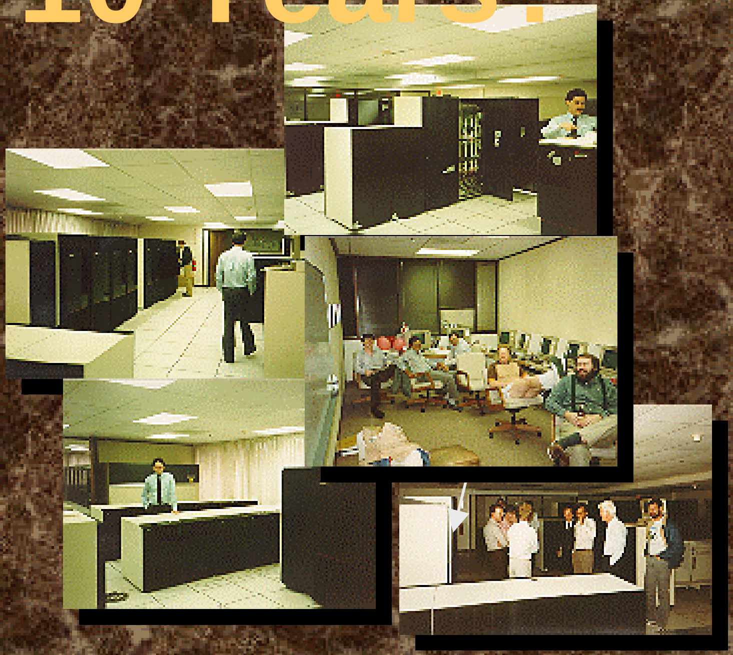
  (1.8 M messages/day)
- And, more coming.....

Windows NT Server
Enterprise Edition



- Scalability
  - 8x SMP support (32x in OEM kit)
  - Larger process memory (3GB Intel)
  - Unlimited Virtual Roots in IIS (web)
- Transactions
  - DCOM transactions (Viper TP mon)
  - Message Queuing (Falcon)
- Availability
  - Clustering (WolfPack)
  - Web, File, Print,DB … servers fail over.

What Happens in 10 Years?



1987: 256 tps

$ 14 million computer

A dozen people

Two rooms of machines

1997: 1,250 tps

$ 50 k$ computer

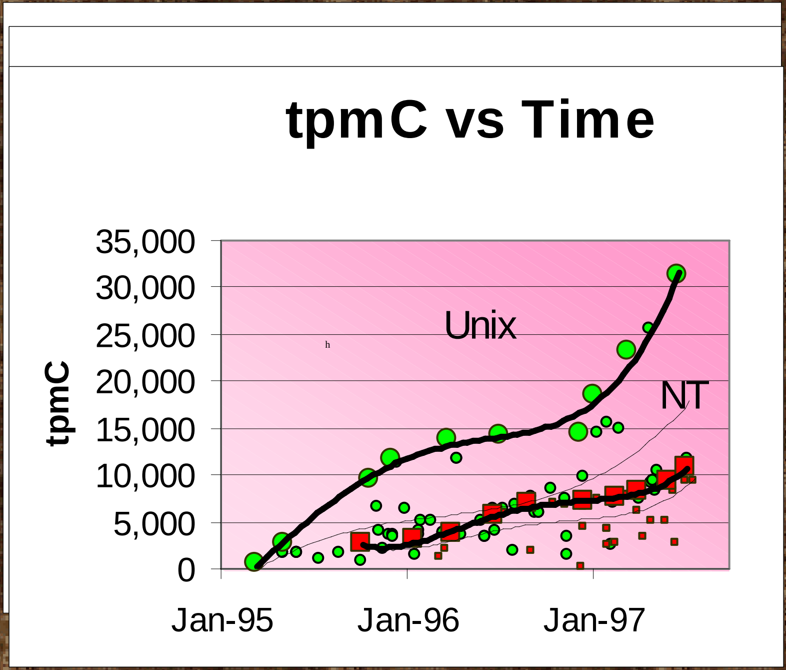
One person

1 micro-dollar per transaction
(1,000x cheaper)

Ready for the next 10 years?

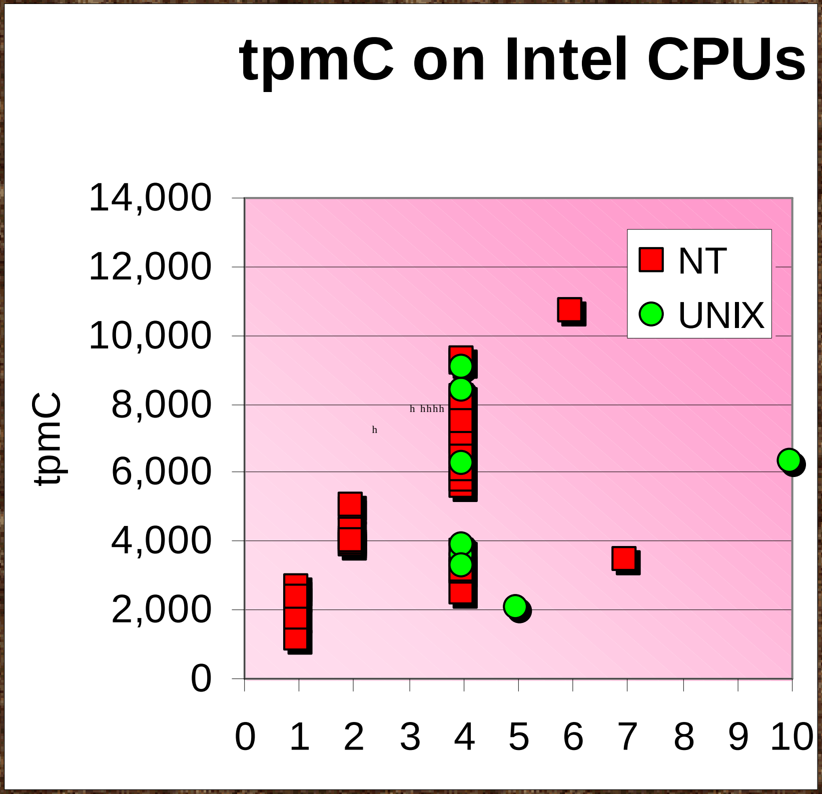
NT vs UNIX SMPs

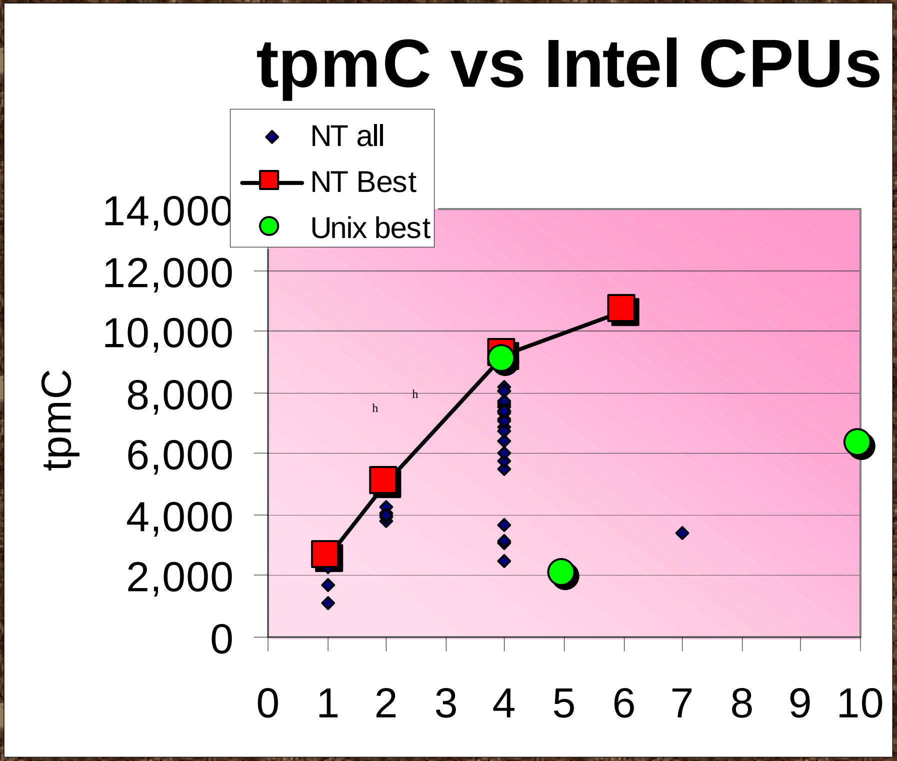
- NT traditionally ran on 1 to 4 cpus
  - Scales near-linear on them
- UNIX boxes: 32-64 way SMPs
  - They do 3x more tpmC
  - They cost 10x more.
- 10 way NT machines are available
  - They cost more
  - They are faster
- My view (shared by many)
  - Need clusters for availability
  - Cluster commodity servers to make huge systems
  - a la Tandem, Teradata, VMScluster, IBM Sysplex, IBM SP2
  - Clusters reduce need for giant SMPs



Transaction Throughput TPC-C

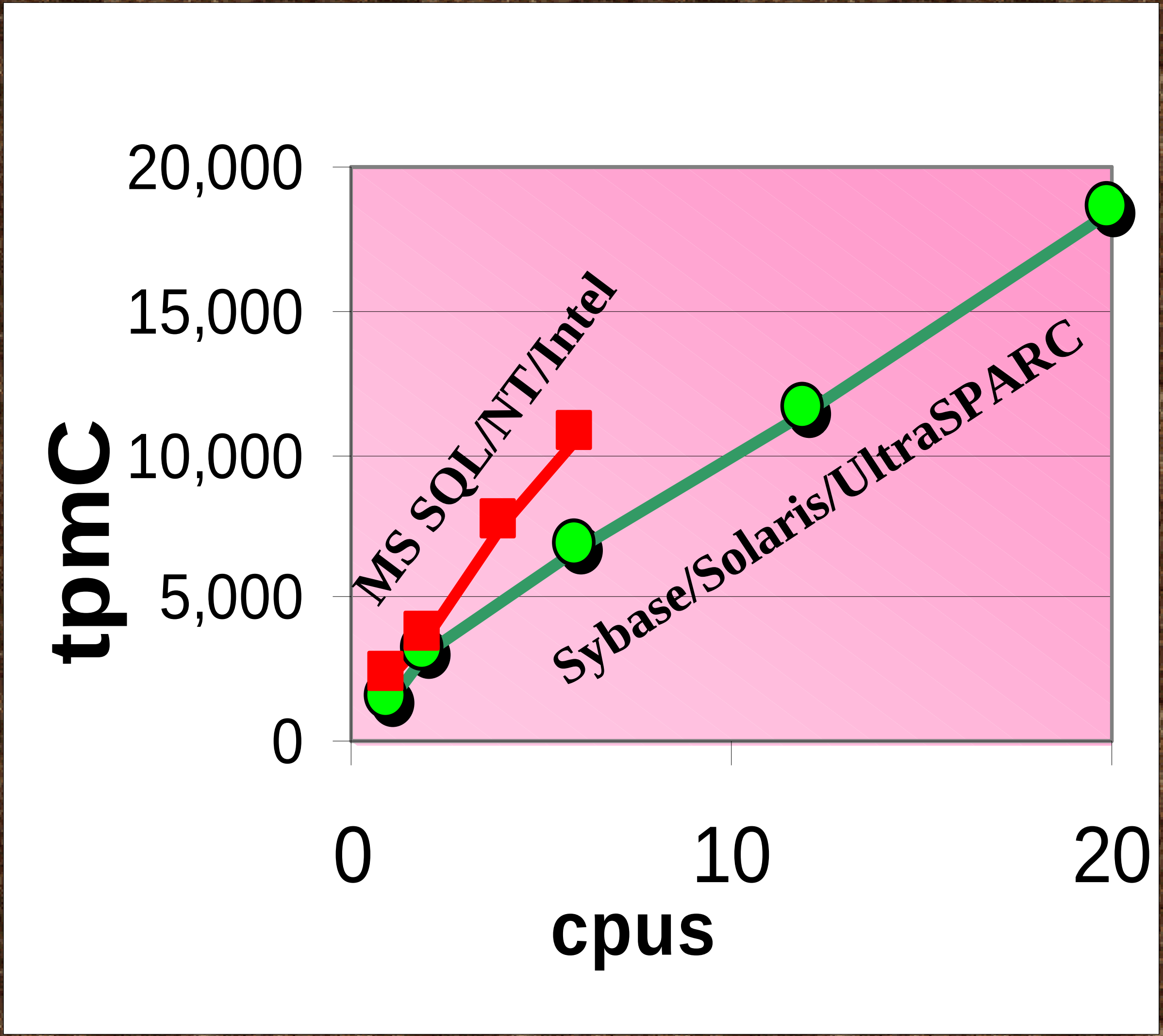
- On comparable hardware: NT scales better!
- SQL Server & NT Improving 250% per year
- NT has best Price Performance (2x cheaper)





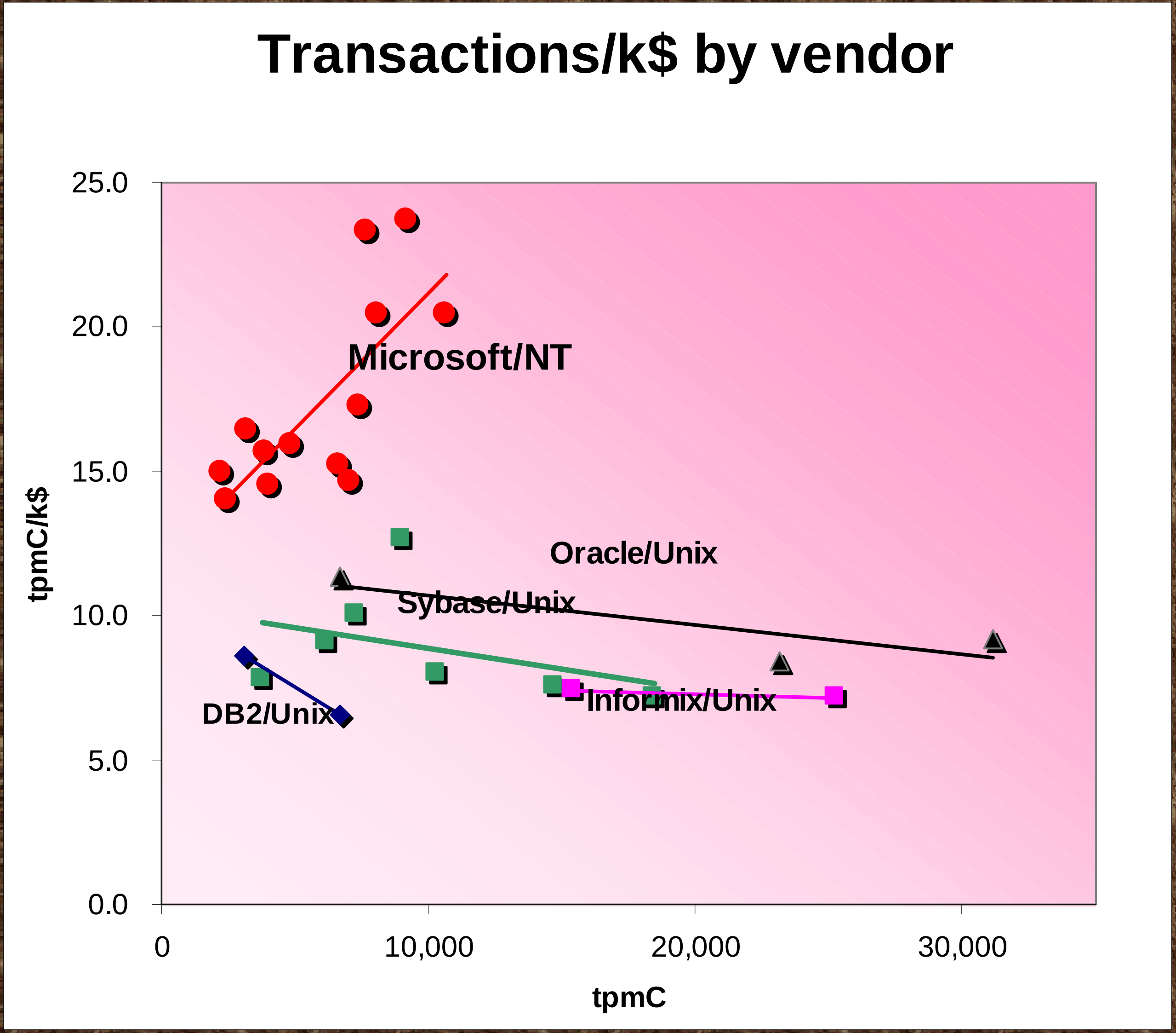
NT Scales Better Than Solaris

- Microsoft SQL NT Intel scales to 6x
- Beats Sybase Solaris UltraSPARC up to 11-way



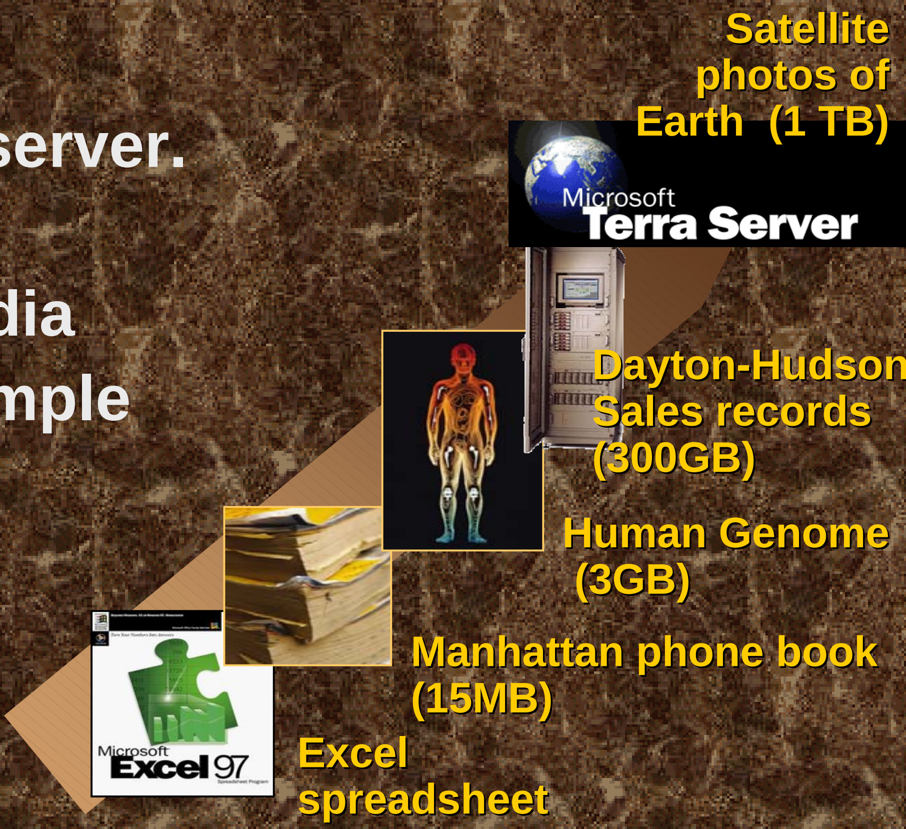
Only NT Has Economy of Scale

- NT is 2x less expensive 40$/tpmC vs 110$/tpmC
- Only NT has economy of scale
- Unix has dis-economy of scale

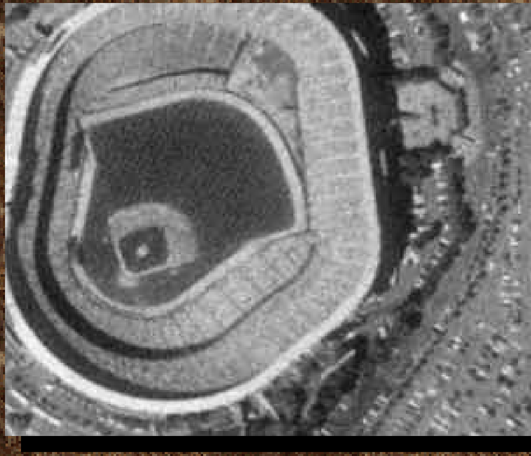


Scaleup To Big Databases?

- NT 4 and SQL Server 6.5
  - DBs up to 1 Billion records,
  - 100 GB
  - Covers most (80%) data warehouses
- SQL Server 7.0
  - Designed for Terabytes
    - Hundreds of disks per server.
    - SMP parallel search
  - Data Mining and Multi-Media
- TerraServer is good MM example


Satellite photos of Earth (1 TB)
Microsoft Terra Server
Dayton-Hudson Sales records (300GB)
Human Genome (3GB)
Manhattan phone book (15MB)
Excel spreadsheet
Microsoft Excel 97

Database Scaleup: TerraServer™



- **Demo NT and SQL Server scalability**
- **Stress test SQL Server 7.0**
- **Requirements**
  - **1 TB**
  - **Unencumbered (put on www)**
  - **Interesting to everyone everywhere**
  - **And not offensive to anyone anywhere**
- Loaded
  - 1.1 M place names from Encarta World Atlas
  - 1 M Sq Km from USGS (1 meter resolution)
  - 2 M Sq Km from Russian Space agency (2 m)
- Will be on web (world’s largest atlas)
- Sell images with commerce server.
- USGS CRDA: 3 TB more coming.

Microsoft® BackOffice™

Microsoft Terra Server

System

- DEC Alpha 4100 (4x smp) +
- 324 StorageWorks Drives (1.4 TB)
- RAID 5 Protected
- SQL Server 7.0
- USGS 1-meter data
  (30% of US)
- Russian Space data
  Two meter resolution images
  (2 M km^2
  2% of earth)

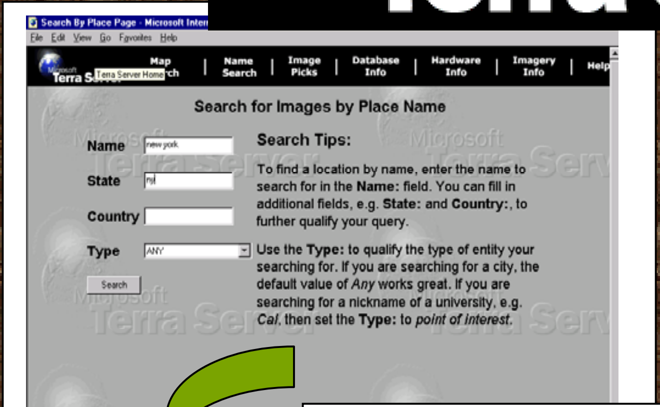
U S G S U.S. Geological Survey National Mapping Information

SPIN-2

digital

Demo

Microsoft Terra Server



Search for Images by Place Name

Name: new york

State: nj

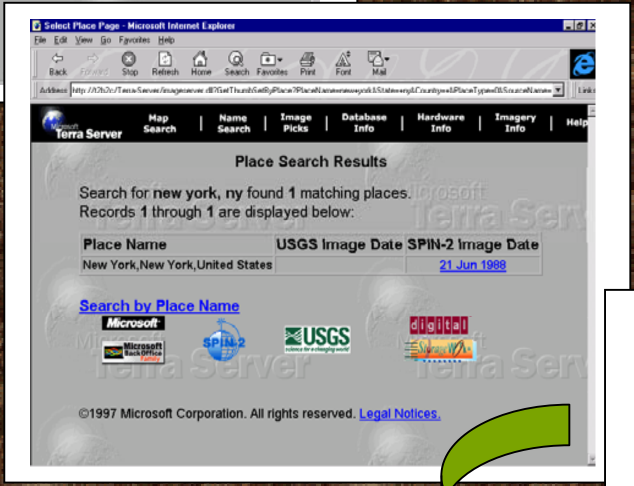
Country:

Type: ANY

Search Tips:

To find a location by name, enter the name to search for in the **Name:** field. You can fill in additional fields, e.g. **State:** and **Country:**, to further qualify your query.

Use the **Type:** to qualify the type of entity your searching for. If you are searching for a city, the default value of *Any* works great. If you are searching for a nickname of a university, e.g. *Cal*, then set the **Type:** to *point of interest*.



Place Search Results

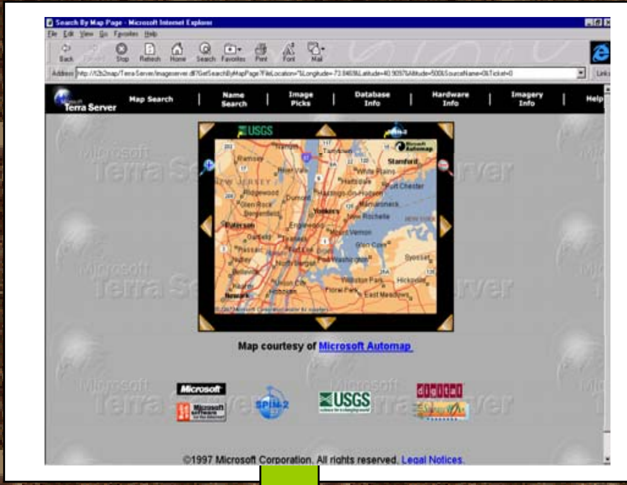
Search for **new york, ny** found **1** matching places.
Records **1** through **1** are displayed below:

Place Name	USGS Image Date	SPIN-2 Image Date
New York,New York,United States		21 Jun 1988

Search by Place Name

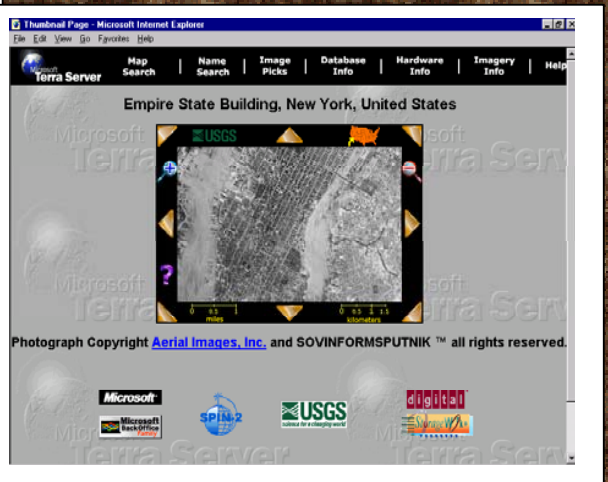
©1997 Microsoft Corporation. All rights reserved. Legal Notices.

Http://t2b2c



Map courtesy of Microsoft Automap.

©1997 Microsoft Corporation. All rights reserved. Legal Notices.



Empire State Building, New York, United States

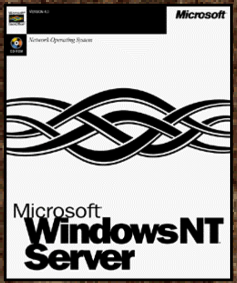
Photograph Copyright Aerial Images, Inc. and SOVINFORMSPUTNIK ™ all rights reserved.

digital™

Manageability

Windows NT 5.0 and Windows 98

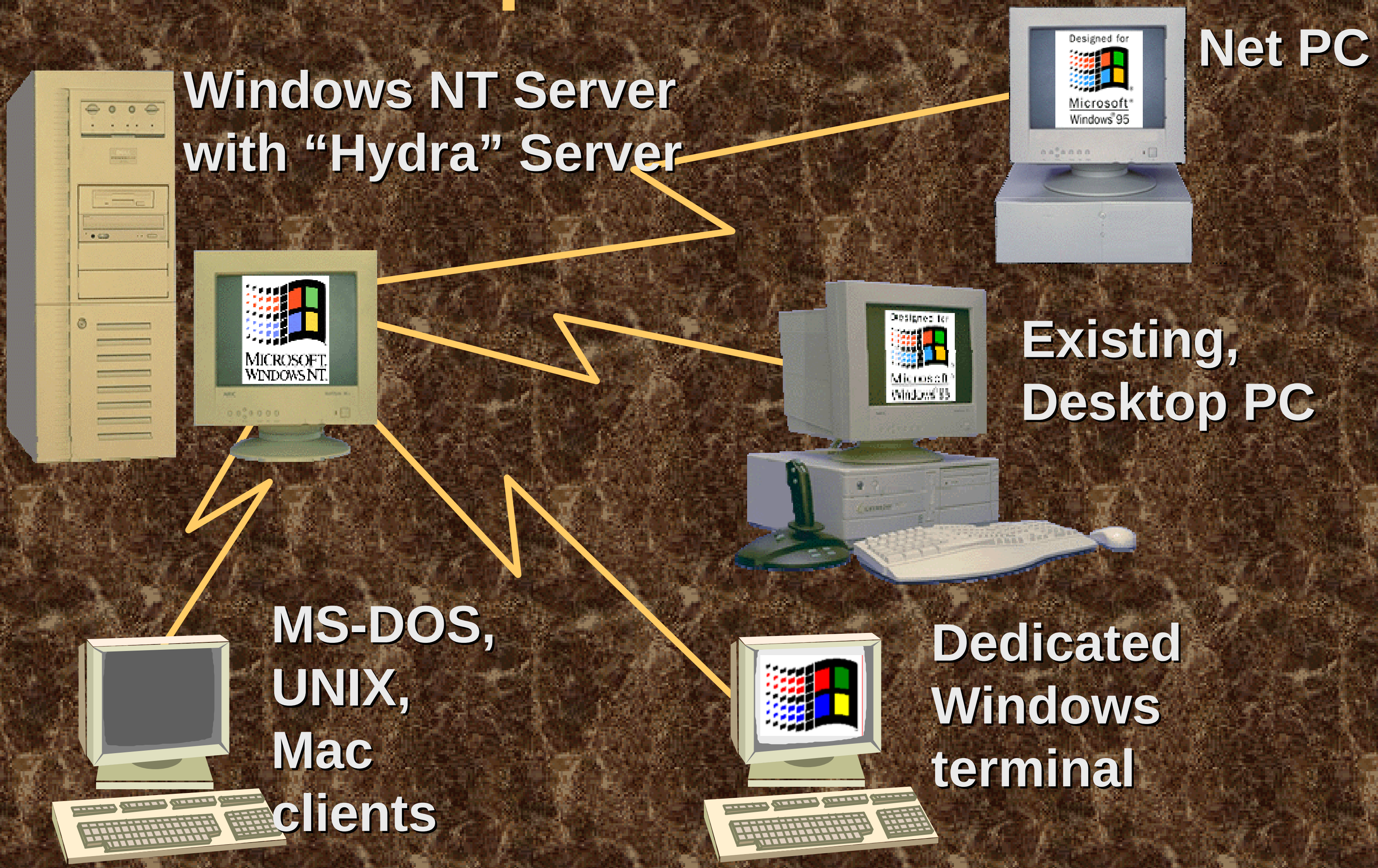
- Active Directory tracks all objects in net
- Integration with IE 4.
  - Web-centric user interface
- Management Console
  - Component architecture
- Zero Admin Kit and Systems Management Server
- PlugNPlay, Instant On, Remote Boot,..
- Hydra and Intelli-Mirroring



Thin Client Support

TSO comes to NT

lower per-client costs


Windows NT Server with “Hydra” Server

Net PC

Existing, Desktop PC

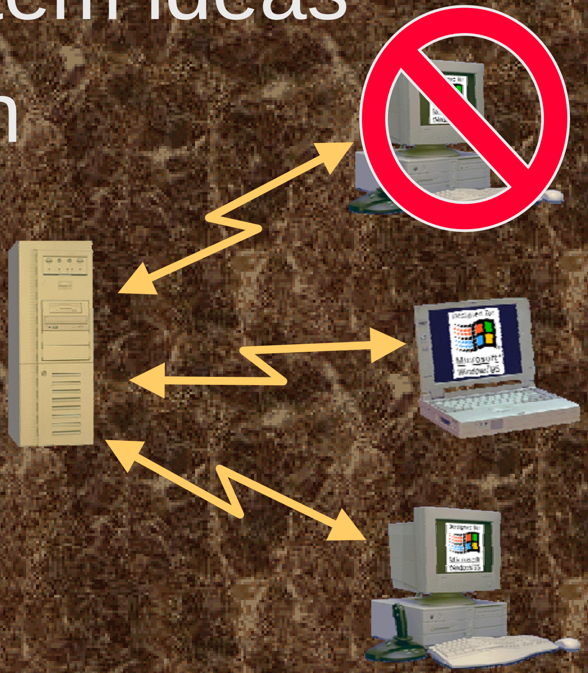
MS-DOS, UNIX, Mac clients

Dedicated Windows terminal

Windows NT 5.0

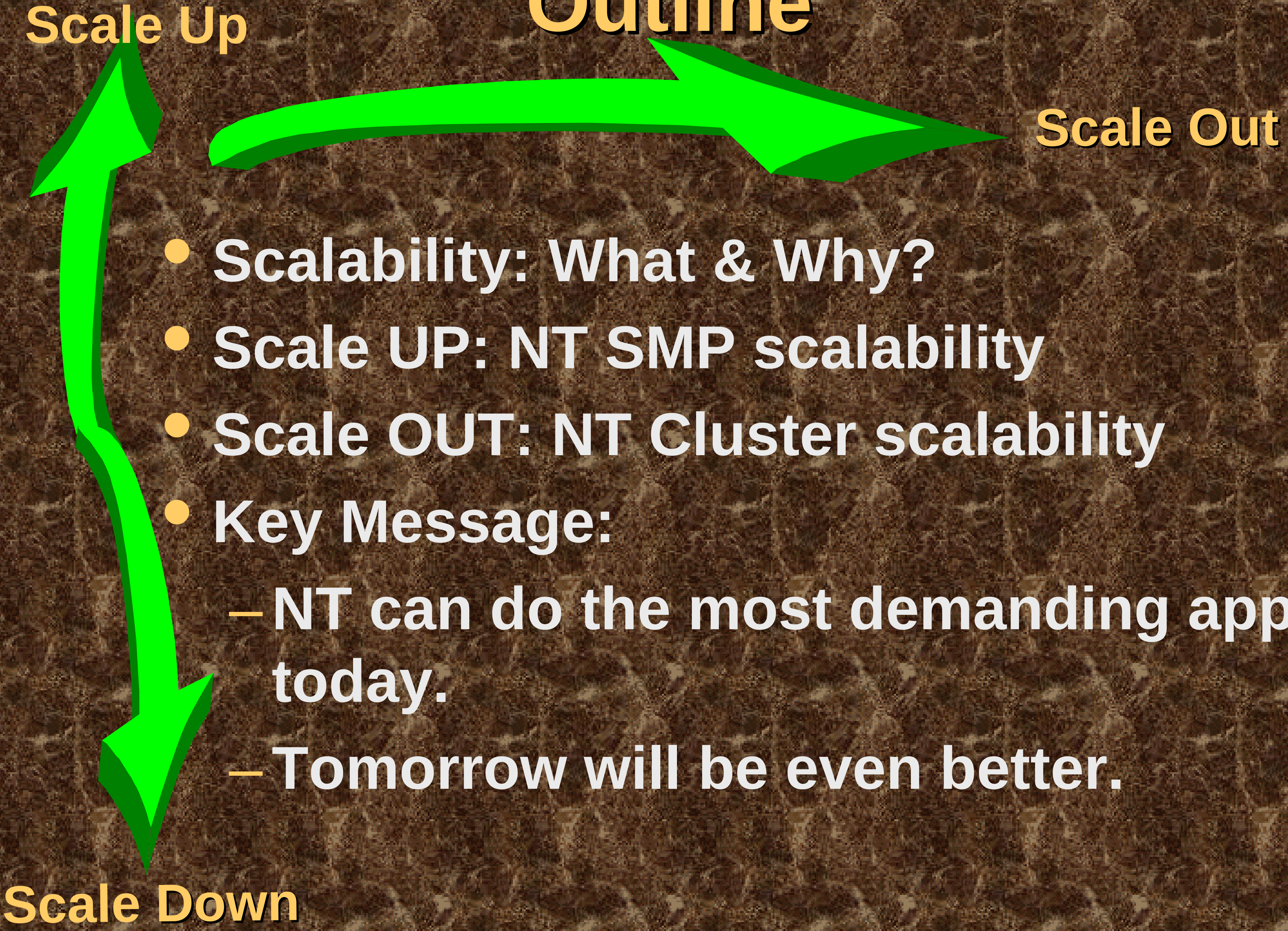
IntelliMirror™

- Extends CMU Coda File System ideas
- Files and settings mirrored on client and server
- Great for disconnected users
- Facilitates roaming
- Easy to replace PCs
- Optimizes network performance



Best of PC and centralized computing advantages

Outline



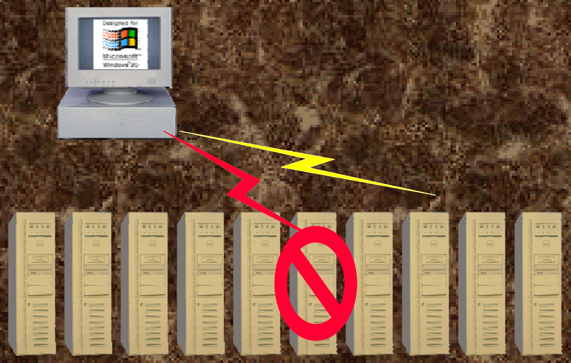
Scale Up

Scale Out

Scale Down

- Scalability: What & Why?
- Scale UP: NT SMP scalability
- Scale OUT: NT Cluster scalability
- Key Message:
  - NT can do the most demanding apps today.
  - Tomorrow will be even better.

Scale OUT
Clusters Have Advantages



- **Fault tolerance:**
  - Spare modules mask failures
- **Modular growth without limits**
  - Grow by adding small modules

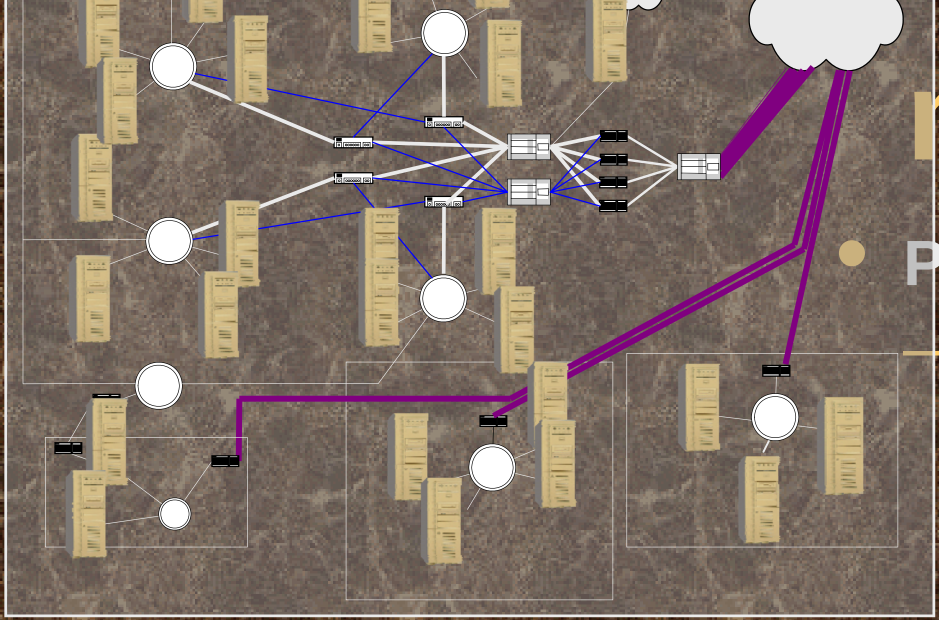
- **Parallel data search**
  - Use multiple processors and disks
- **Clients and servers made from the same stuff**
  - Inexpensive: built with commodity CyberBricks

How scaleable is NT?? The Cluster Story

- 16-node Tandem Cluster
  - 64 cpus
  - 2 TB of disk
  - Decision support
- 45-node Compaq Cluster
  - 140 cpus
  - 14 GB DRAM
  - 4 TB RAID disk
  - OLTP (Debit Credit)
    - 1 B tpd (14 k tps)

TANDEM

COMPAQ



microsoft.com

- **Production**
  - Windows NT.4 and IIS.3
    - 20 HTTP,
    - 3 download,
    - 3 FTP
    - 5 SQL 6.5
    - Index Server + 3 search
- **Stagers**
  - Site Server for content
  - DCOM Publishing wizard
- **Network**
  - 6 DS3
  - 4 TB/day download capacity
- Replicas in UK and Japan

- 90m hits/day
  - 17m page views
  - #4 site on Internet
- 900k visitors per day
- Not cheap
  - Data Centers
  - Bandwidth
  - 27 people on content
  - 22 people on systems

Tandem 2 Ton

- 2 TB SQL database
- 1.2 TB user data
- 16 node cluster
- 64 cpus, 480 disks
- Decision support
  parallel data-mining

- Will be Wolf Pack aware
- Demoed at DB Expo in
- ServerNet™ interconnect

TANDEM

Billion Transactions per Day Project

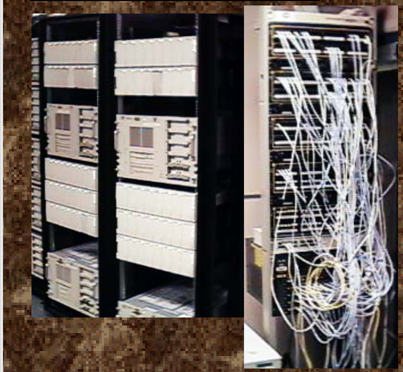
- Built a 45-node Windows NT Cluster
  (with help from Intel & Compaq)
  > 900 disks
- All off-the-shelf parts
- Using SQL Server &
  DTC distributed transactions
  DCOM & ODBC clients
  on 20 front-end nodes
- DebitCredit Transaction
- Each server node has 1/20 th of the DB
- Each server node does 1/20 th of the work
- 15% of the transactions are “distributed”

COMPAQ

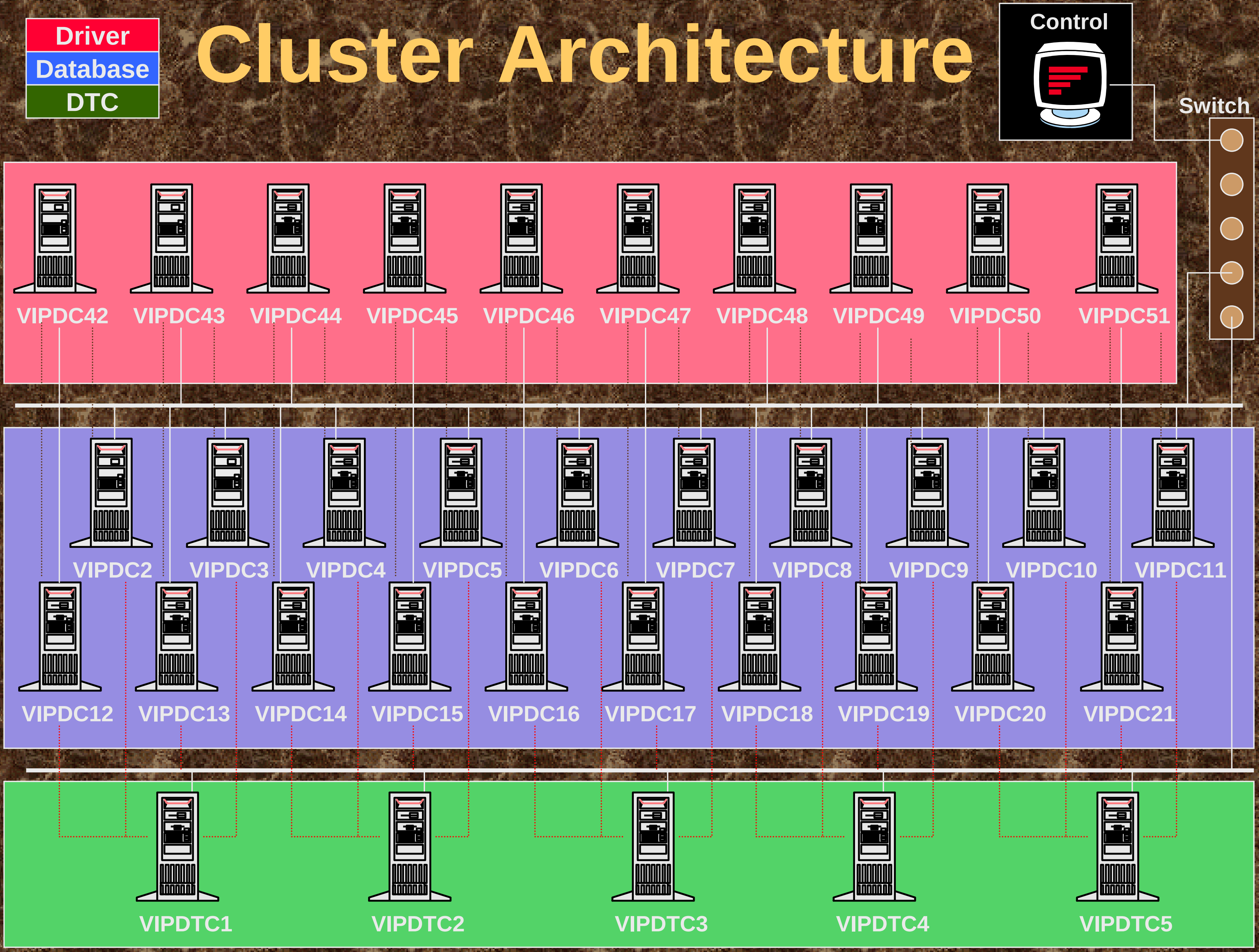
Billion Transactions Per Day Hardware

- 45 nodes (Compaq Proliant)
- Clustered with 100 Mbps Switched Ethernet
- 140 cpu, 13 GB, 3 TB (RAID 1, 5).

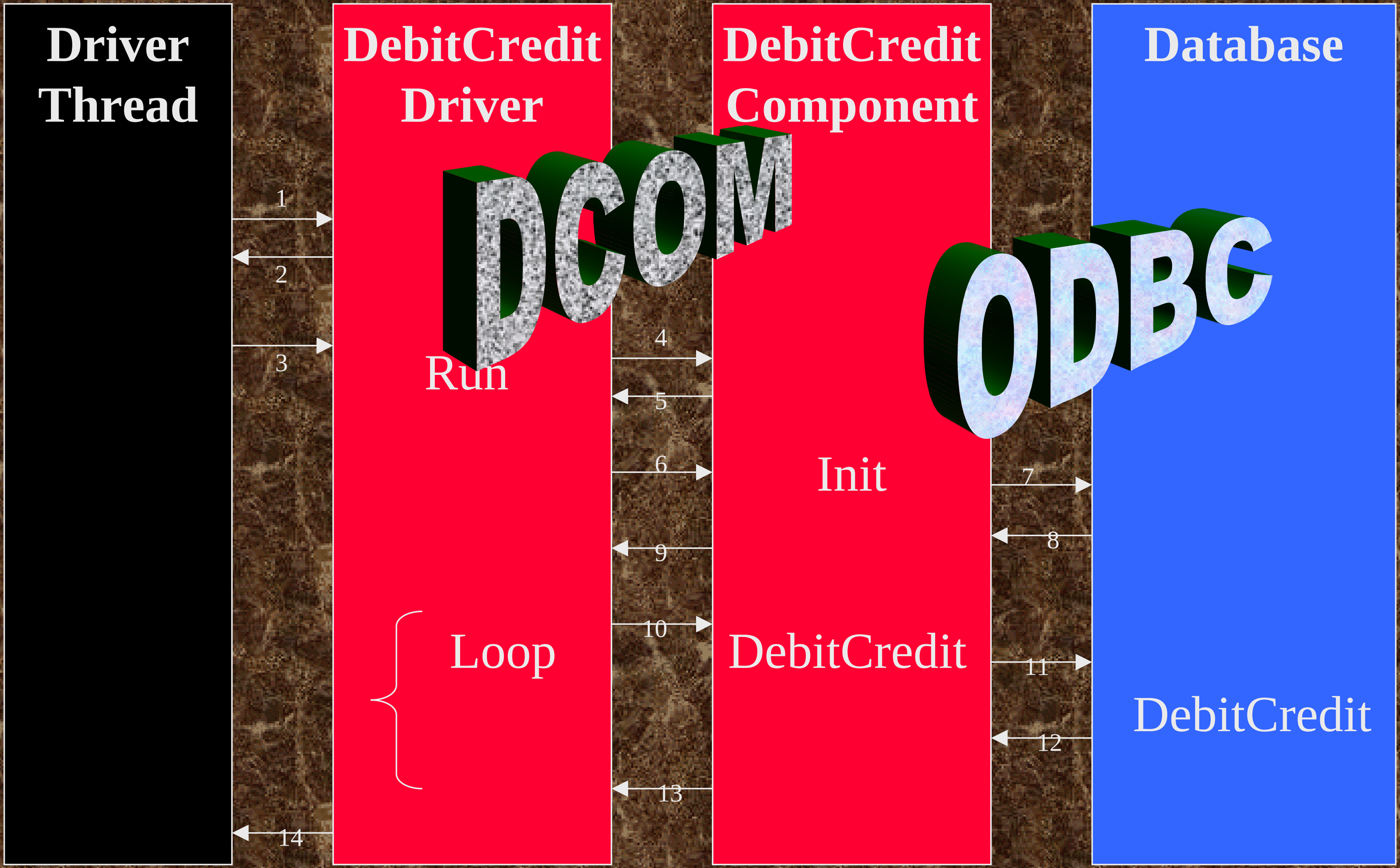
Type	nodes	CPUs	DRAM	ctlrs	disks	RAID space
Workflow MTS	20 Compaq Proliant 2500	20x 2	20x 128	20x 1	20x 1	20x 2 GB
SQL Server	20 Compaq Proliant 5000	20x 4	20x 512	20x 4	20x 36x4.2GB 7x9.1GB	20x 130 GB
Distributed Transaction Coordinator	5 Compaq Proliant 5000	5x 4	5x 256	5x 1	5x 3	5x 8 GB
TOTAL	**45**	**140**	**13 GB**	**105**	**895**	**3 TB**



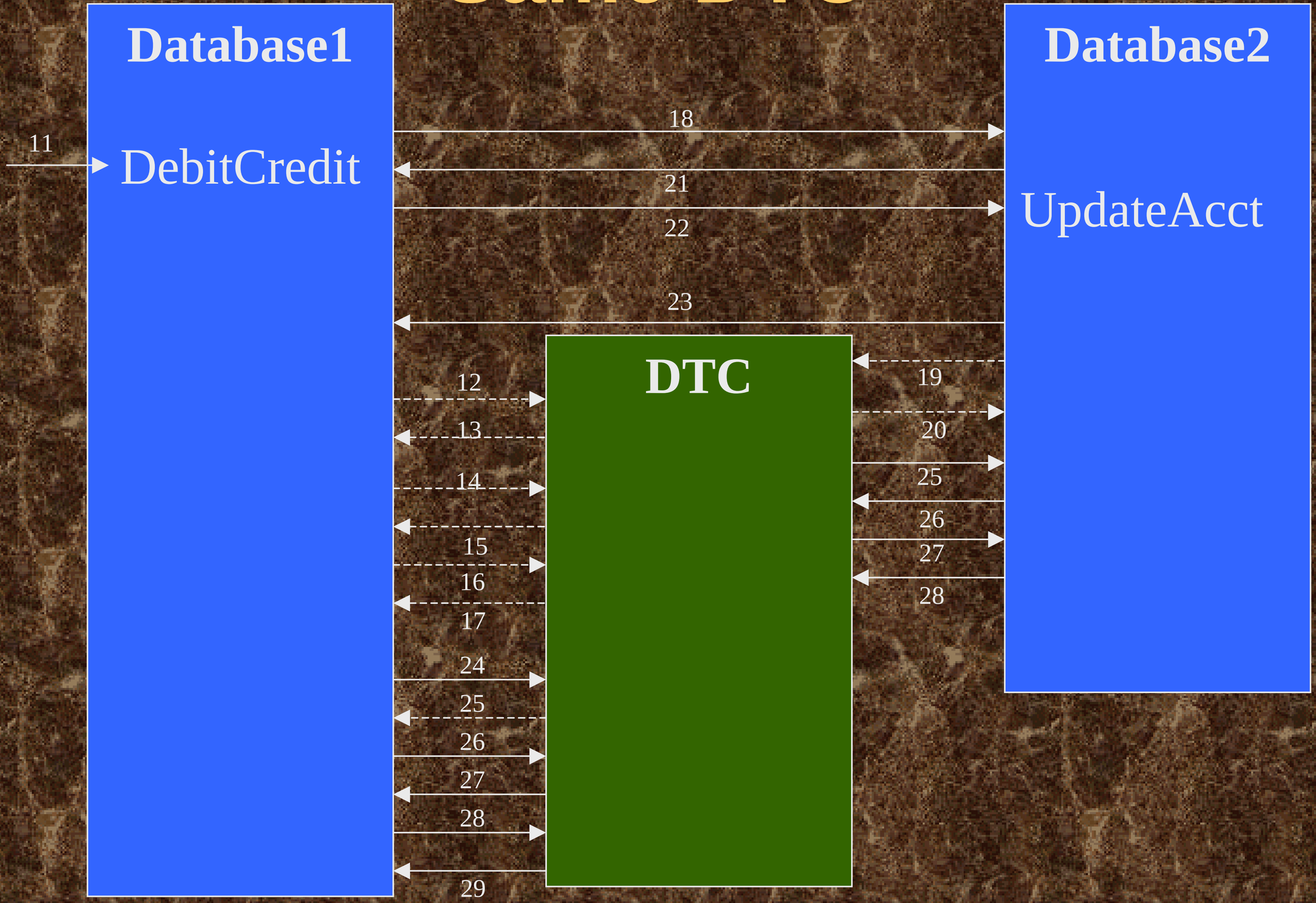
Cluster Architecture



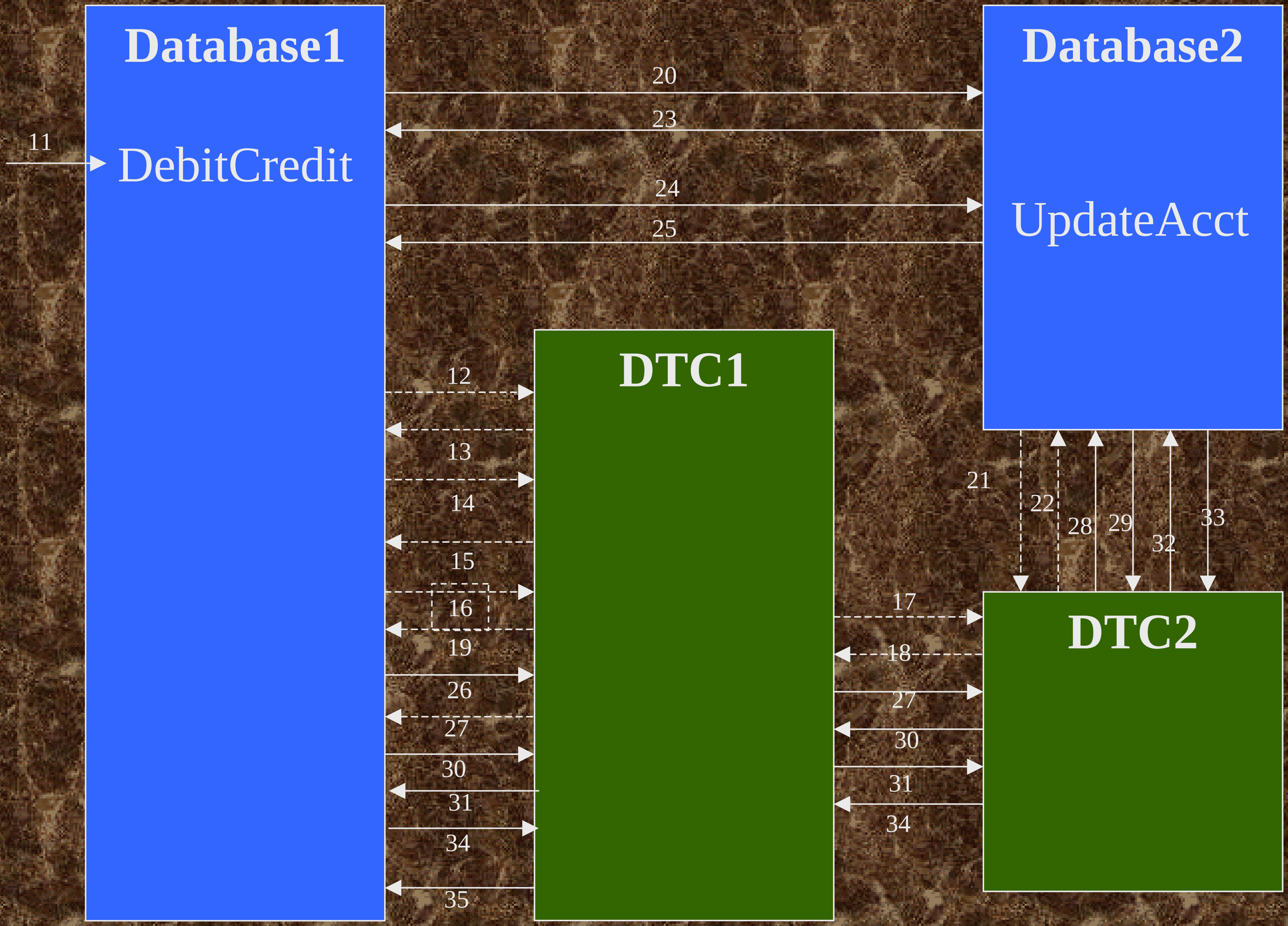
Local Debit Credit



Distributed Debit Credit - Same DTC

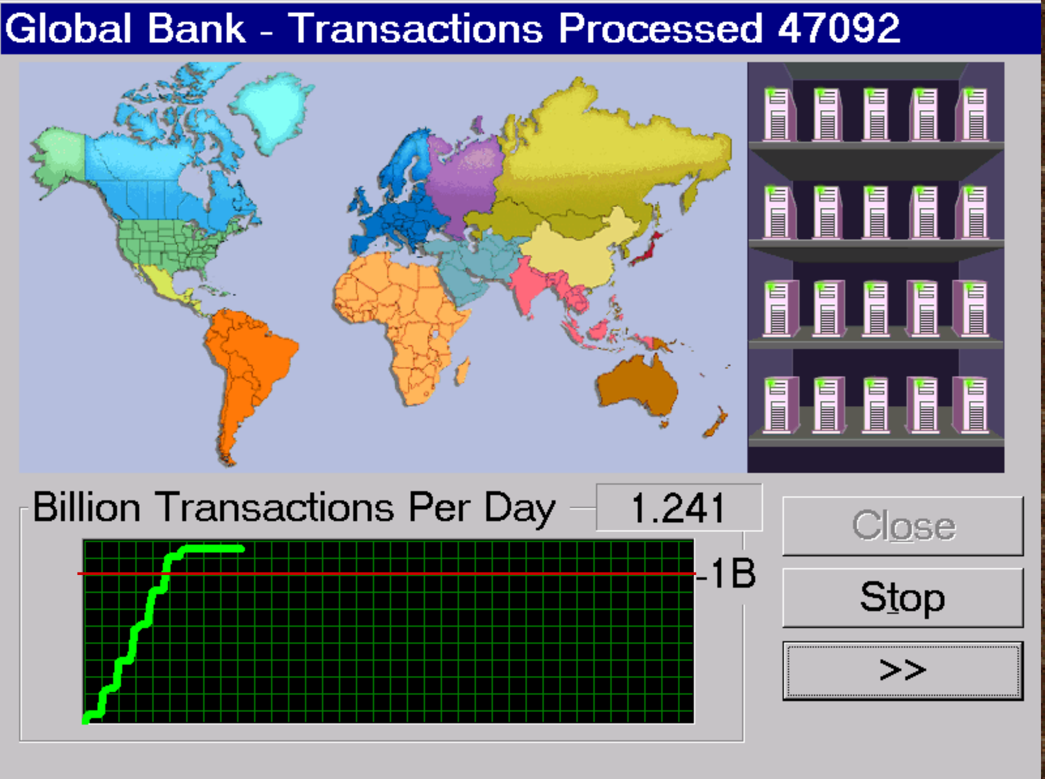


Distributed Debit Credit - Different DTC



1.2 B tpd

- 1 B tpd ran for 24 hrs.
- Out-of-the-box software
- Off-the-shelf hardware
- AMAZING!



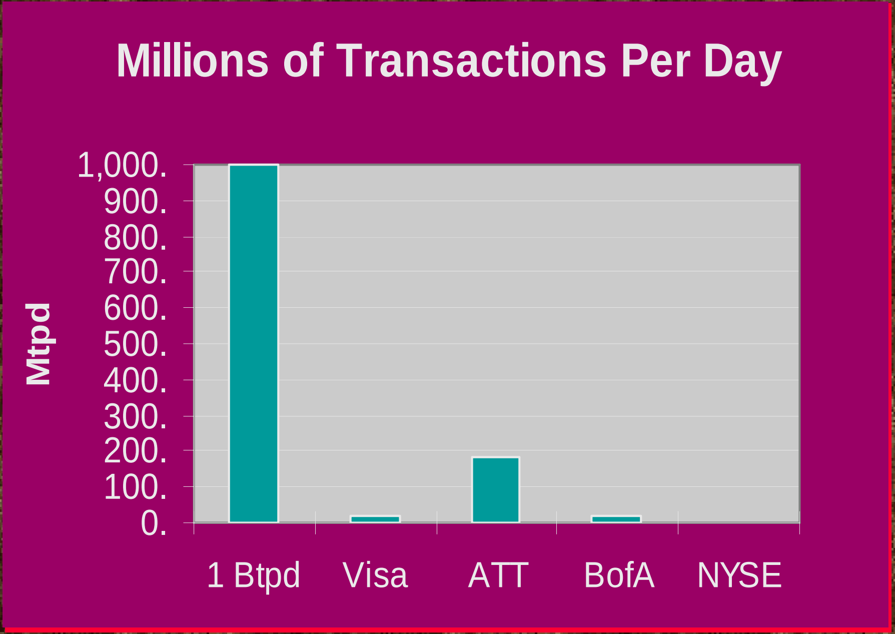
Global Bank - Transactions Processed 47092

Billion Transactions Per Day 1.241

- Sized for 30 days
- Linear growth
- 5 micro-dollars per transaction

How Much Is 1 Billion Tpd?

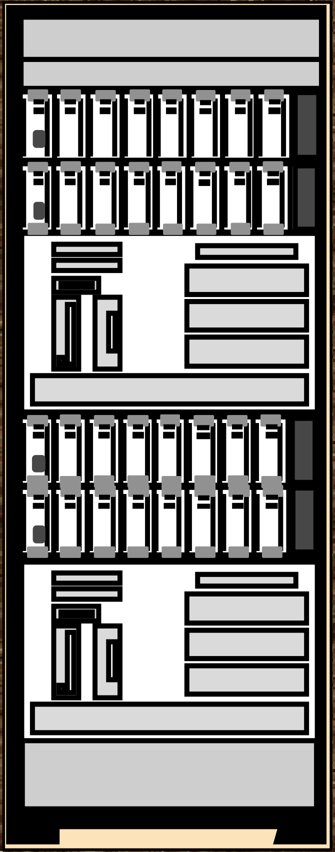
- 1 billion tpd = 11,574 tps
  ~ 700,000 tpm (transactions/minute)
- ATT
  - 185 million calls per peak day (worldwide)
- Visa ~20 million tpd
  - 400 million customers
  - 250K ATMs worldwide
  - 7 billion transactions (card+cheque) in 1994
- New York Stock Exchange
  - 600,000 tpd
- Bank of America
  - 20 million tpd checks cleared (more than any other bank)
  - 1.4 million tpd ATM transactions
- Worldwide Airlines Reservations: 250 Mtpd



1 B tpd: So What?

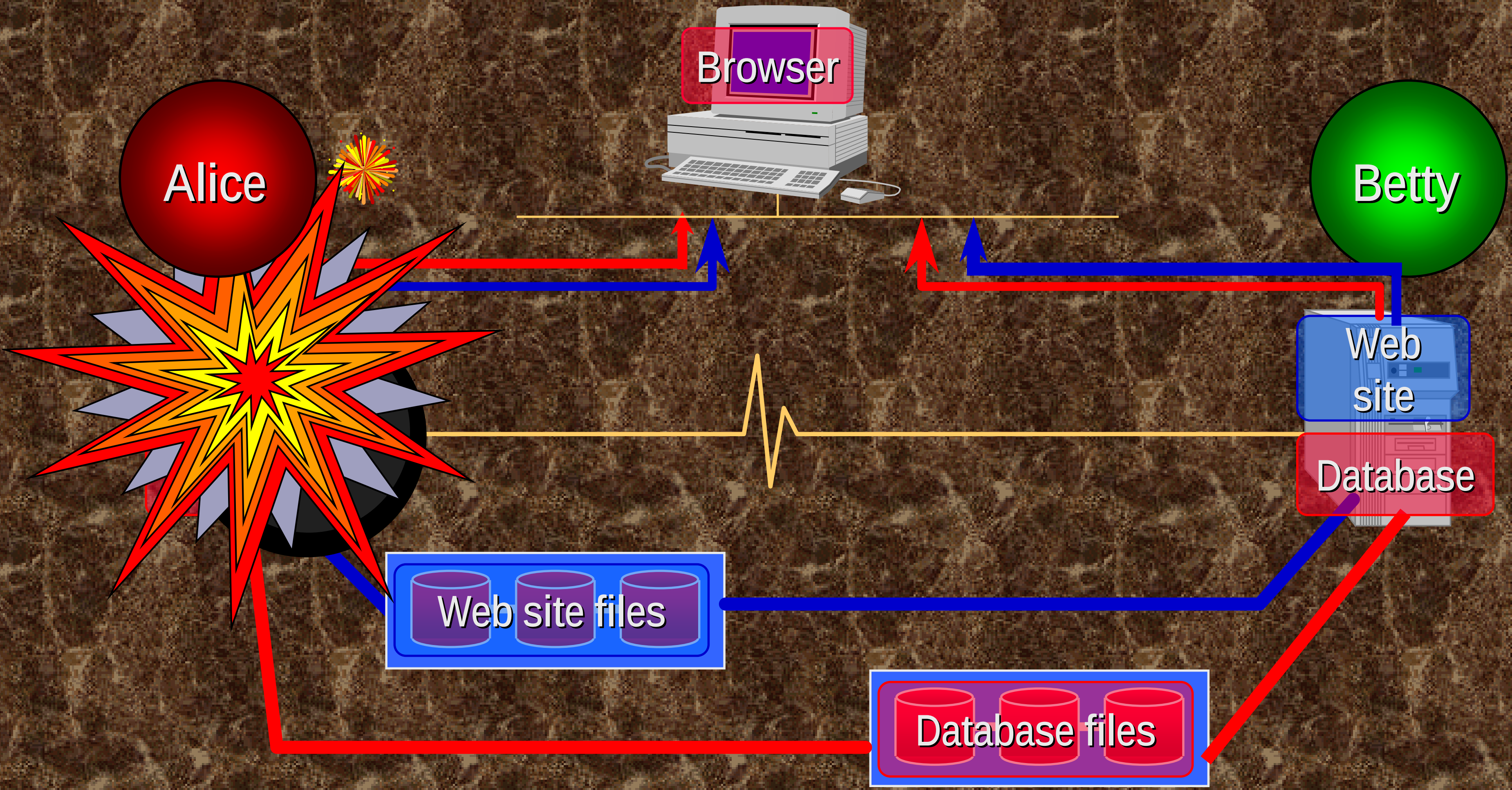
- Shows what is possible, easy to build
  - Grows without limits
- Shows scaleup of DTC, MTS, SQL…
- Shows (again) that shared-nothing clusters scale
- Next task: make it easy.
  - auto partition data
  - auto partition application
  - auto manage & operate

Cluster Server: High Availability

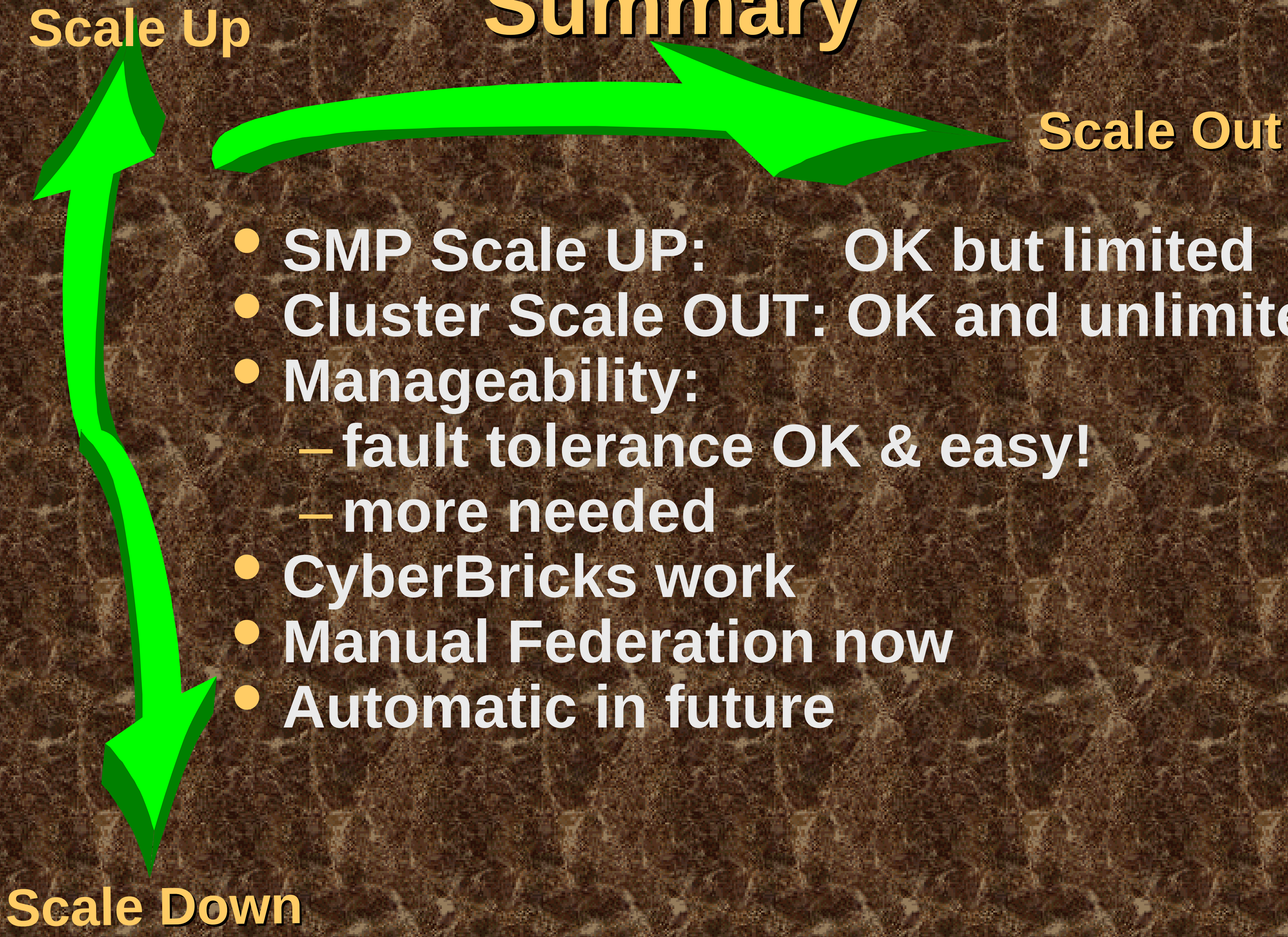


- Multiple servers form one system
- Industry standard APIs and hardware
- Server application and tools support
  - IIS web server
  - File and Print servers
  - IP and NetName failover
  - Transaction and Queue Server failover
  - SQL Server, Enterprise edition
- Tight integration with Windows NT -- its easy!
- Two-Node clusters now (2 to 20 cpus)
- 16 node soon (2 to 192 cpus).

WolfPack Cluster
IIS & SQL Failover Demo



Summary



- SMP Scale UP: OK but limited
- Cluster Scale OUT: OK and unlimited
- Manageability:
  - fault tolerance OK & easy!
  - more needed
- CyberBricks work
- Manual Federation now
- Automatic in future

Scalability Research Problems

- Automatic everything
- Scaleable applications
  - Parallel programming with clusters
  - Harvesting cluster resources
- Data and process placement
  - auto load balance
  - dealing with scale (thousands of nodes)
- High-performance DCOM
  - active messages meet ORBs?
- Process pairs, other FT concepts?
- Real time: instant failover
- Geographic (WAN) failover