Windows NT Scalability

Jim Gray
Microsoft Research
Gray@Microsoft.com
• 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?

• 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

• 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 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
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.
• 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 km² 2% of earth)
Microsoft Terra Server

Search for images by Place Name

Search Tips:
- 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 name of a university, e.g. Cal, then set the Type: to point of interest.

http://t2b2c
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

Existing, Desktop PC

Net 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

• 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)
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
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”
**Billion Transactions Per Day Hardware**

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

<table>
<thead>
<tr>
<th>Type</th>
<th>nodes</th>
<th>CPUs</th>
<th>DRAM</th>
<th>ctrlrs</th>
<th>disks</th>
<th>RAID space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflow MTS</td>
<td>20</td>
<td>20x</td>
<td>20x</td>
<td>20x</td>
<td>20x</td>
<td>20x</td>
</tr>
<tr>
<td>Compaq Proliant 2500</td>
<td></td>
<td>2</td>
<td>128</td>
<td>1</td>
<td>1</td>
<td>2 GB</td>
</tr>
<tr>
<td>SQL Server</td>
<td>20</td>
<td>20x</td>
<td>20x</td>
<td>20x</td>
<td>20x</td>
<td>20x</td>
</tr>
<tr>
<td>Compaq Proliant 5000</td>
<td></td>
<td>4</td>
<td>512</td>
<td>4</td>
<td></td>
<td>20x</td>
</tr>
<tr>
<td></td>
<td>36x4.2GB</td>
<td>7x9.1GB</td>
<td></td>
<td></td>
<td></td>
<td>130 GB</td>
</tr>
<tr>
<td>Distributed Transaction Coordinator</td>
<td>5</td>
<td>5x</td>
<td>5x</td>
<td>5x</td>
<td>5x</td>
<td>5x</td>
</tr>
<tr>
<td>Compaq Proliant 5000</td>
<td></td>
<td>4</td>
<td>256</td>
<td>1</td>
<td>3</td>
<td>8 GB</td>
</tr>
<tr>
<td>TOTAL</td>
<td>45</td>
<td>140</td>
<td>13 GB</td>
<td>105</td>
<td>895</td>
<td>3 TB</td>
</tr>
</tbody>
</table>

TOTAL 45 140 13 GB 105 895 3 TB
Cluster Architecture

VIPDC42  VIPDC43  VIPDC44  VIPDC45  VIPDC46  VIPDC47  VIPDC48  VIPDC49  VIPDC50  VIPDC51

VIPDC2   VIPDC3   VIPDC4   VIPDC5   VIPDC6   VIPDC7   VIPDC8   VIPDC9   VIPDC10  VIPDC11

VIPDC12  VIPDC13  VIPDC14  VIPDC15  VIPDC16  VIPDC17  VIPDC18  VIPDC19  VIPDC20  VIPDC21

VIPDTC1  VIPDTC2  VIPDTC3  VIPDTC4  VIPDTC5
Local Debit Credit

Driver Thread

DebitCredit

DebitCredit Component

Database

DCOM

Run

Loop

Init

DebitCredit

DebitCredit

DebitCredit
Distributed Debit Credit - Same DTC
Distributed Debit Credit - Different DTC

- Database1
- DebitCredit
- DTC1
- Database2
- UpdateAcct
- DTC2
1.2 B tpd

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

- 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
1B 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