How Many Linux Systems Can Dance on a Single S/390?

David Boyes
Sine Nomine Associates (SIN)
Session 5538/9202

Agenda
- Overview of ISP/IDC Environment
- Construction of the Test Case
- Process of Testing
- What We Learned
- So What’s the Point Anyway?
- Areas for Further Research

Questions
- Please hold questions until the end -- I’ve got lots to talk about, and I want to make sure we get through all of it.
How Many Linux Systems Can Dance on a Single S/390?

Overview of ISP/IDC Requirements

Internet Service Providers (ISP)s and Internet-oriented Data Centers (IDCs) have similar requirements:
- standard open-source applications (sendmail, bind, UCB POP3, UW IMAP, WUFTPD, INN, etc)
- primarily Unix-based environment
- IP-centric (some Novell, some NETBIOS)

Overview of ISP/IDC Requirements

Primary differentiator is scalability and TCO:
- IDC requires substantially larger scalability (avg 5000+ systems for industrial scale)
- Target TCO computation for traditional solution: $1500/sq ft/month
  - total operational cost, including staff, environments, operation and management software, etc.

Network Cost of Ownership

Source: The Registry
Overview of ISP/IDC Requirements

- Secondary differentiator is time to market (TTM):
  - avg for discrete machines = 7 days from payment to delivery
  - high-volume sources (Exodus, AboveNet) avg 4-5 days to delivery
- Most business ISP/IDC customers expect dedicated servers to guarantee SLAs.

Horizontal Vs Vertical Scaling

- Horizontal:
  - Well suited to distributed apps and client/server
  - Use of load balancing hardware hides complexity

- Vertical:
  - Well suited to interactive user sessions and applications
  - Complex to integrate due to conflicting application requirements
"Well, this is a pretty mess you’ve gotten us into…"

Customer looking at requirements for infrastructure buildout for managed router services:
- 250 initial customers
- DNS and Usenet News/INN only for first service offering (later offerings based on success of managed router service)

System Count: Discrete Solution
- estimating 2 Sun UE2 class systems for DNS; 1 Sun UE1000 system for INN due to I/O requirements.
  - System requirement replicated for each customer.
  - Implies 2 RU per UE2; 4 RU per UE1000 + disk array (2-4 RU)
- 3 systems per customer: 750 machines!

Support Infrastructure: Discrete Solution
- 3 physical LAN ports
- 1/3 of a rack
- VLAN configuration
- cabling and cabling management
- IP address allocation & routing policy
- Tivoli management agent license
- Tivoli TSM backup client license
- etc, etc, etc
How Many Linux Systems Can Dance on a Single S/390?

The Approach
- Customer unwilling to commit without proof of concept.
- Customer uncomfortable with “bucking the trend” and concerned about perception of S/390 vs traditional solution.
- Solution: do a study and push the technology hard to determine feasibility!

Objects of Study
- Scalability of Linux on System/390
- Compatibility and Applications Support
- Suitability of Linux on System/390 for ISP/IDC server platform
- Just plain curiosity

Architecture of Study
- Must resemble a “real” application prevalent in an ISP/IDC/ASP environment.
- Must show:
  - traditional ISP applications (DNS, News, NFS, WWW server)
  - integration of system management and connectivity management
  - viability of virtual server and risk.
How Many Linux Systems Can Dance on a Single S/390?

Test Plan Able/Baker

- Small scale tests (250, 2750, 10000 images)
- Relied on test scripting and easy source portability.
- Determined that all-out testing was required.

Test Plan: Charlie

**TEST PLAN: Charlie**

**Configuration Layout**

- Small Scale Tests
  - 250, 2750, 10000 images

**Test Driver Systems**

- L/390 images
  - 1 virtual router system per 200 interfaces

**Data Cache Systems**

- L/390 images
  - 1 data server image per 1000 client systems

**WWW Server Instances**

- L/390 images
  - 38,000 images (Goal: 99,999 images)

**Configuration Instance**

- Shared /usr and /bin

**Test Driver Systems**

- L/390 images
  - 1500+ images

**R/O mini-disk link**

- TCP over IUCV
- NFS over TCP
- HTTP over TCP

Lessons Learned

- Substantial operational advantages accrue from SCIF common console and VM system resource instrumentation and management.
  - Increased security and system resource monitoring
  - I/O modeling information
Lessons Learned

Default Linux idle task management concept is not well-suited for hypervisor environments.
- Default 100 hz timer pops consume substantial resources for no benefit if system is idle.
- Must be adjusted proportionately -- other important timing functions are derived from this value.

Lessons Learned

Linux for S/390 reacts proportionally to resource constraints.
- SLA management can be reported and managed via VM resource controls for single-application Linux instances.
- Further experimentation seems to indicate that limiting Linux paging by using large virtual machines is advantageous for large farms (allows VM to make more intelligent resource mgmt decisions).

Lessons Learned

VM is critical to large scale Linux for System/390 scalability.
- 15 LPARs and/or limited management in VIF do not offer sufficient cost/benefit to make the case for Linux on S/390 iron.
- Loss of VM resource management and error recovery substantially complicates system management.
How Many Linux Systems Can Dance on a Single S/390?

Lessons Learned

◆ Applications are directly source-compatible between Intel-based Linux and S/390-based Linux where supporting devices exist.
  - Compute-intensive apps work, but may not be optimum for S/390 unless interacting with other S/390 resources (eg, DB/2, etc).
  - Use of IEEE HW FP is significant (20-30% faster than emulation code depending on problem and instruction mix).

Lessons Learned

◆ A measure of high availability is inherited from the S/390 HW.
◆ Software HA is still somewhat limited and requires significant planning:
  - multiple network stacks
  - dynamic routing
  - service failover during CPU PM

Customer Outcome

◆ Customer is now creating between 15 and 30 virtual systems per day on a new 9672.
◆ Clients of the service are pleased with the uptime and low cost.
◆ Virtual system deployment completely automated (integrated into WWW front-end and back-end business systems).
**How Many Linux Systems Can Dance on a Single S/390?**

**Why?**
- TCO for traditional solution: $1500/sq foot/month.
- Averages:
  - 3500-7000 discrete systems
  - 15,000-20,000 square feet
  - 3500-7000 network cables and LAN ports at $150/port
  - 3500-7000 power cables
  - Time to market: 4-7 days

**Why Not!**
- 1 to 41,000+ systems: 400 square ft (G5+Shark/EMC cabinet + misc routers)
- Time to Market: about 90 seconds per virtual machine created
- 2 high-capacity network cables (DS3/OC3/OC12 plus ESCON cabling to Cisco 7xxx+CIPs)
- 2 power cable per cabinet.

**Simplicity!**

**Where to Go?**
- Test Plan Omega: 100,000 images.
- Multi-physical box clustering
- Global clusters
- Non-S/390 Virtual Machines
- Power Consolidation
How Many Linux Systems Can Dance on a Single S/390?

**Test Plan Omega**

- Push a single S/390 system to the limit: 100,000 systems
  - Successfully deployed 97,943 images on a dedicated 9672-ZZ7 with VM

**Multi-CPU Clusters**

- Use CSE or ISFC to build linked physical clusters (TSAF limits size of cluster to 8).
- Separate applications from network processing/allow PM of individual CPUs w/o interrupting service to entire complex.
- See earlier notes wrt to high-availability planning -- critical to this effort.
- WORKS TODAY WITH VM!

**Global CPU Clusters**

- Link physical systems over long distances
- Value: global companies, large WWW hosting facilities with replication between centers.
- Tested between Virginia and CEDEC, France (network in US, apps in Europe)
How Many Linux Systems Can Dance on a Single S/390?

Non-S/390 Virtual Machines

- Why should VM emulate only the S/390 architecture?
- Can be done acceptably today with Linux for S/390 for almost any popular micro architecture:
  - Intel 486/Pentium
    (good enough to run NT Server!)
  - Macintosh
  - Apple II
  - Commodore 64 (I’m NOT kidding!)

Non-S/390 Virtual Machines

- Hand optimization of code will address speed concerns
- Recoding portions of bochs in 390 BAL dramatically increased speed:
  - Original code: P70-class
  - Assembler version: P200-class
- Rapid deployment of NT images is straightforward for fixed offerings

Power Consolidation

- Interesting development is interest in server consolidation due to recent power problems in California and other locations.
### Power Comparison

<table>
<thead>
<tr>
<th>Discrete Servers (1000 Systems)</th>
<th>G6-based Farm (1000 images)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 18 KW CPU (Sun U2)</td>
<td>- 6.1KW CPU</td>
</tr>
<tr>
<td>- 5 KW (Cisco Routers)</td>
<td>- 5 KW (Cisco Routers)</td>
</tr>
<tr>
<td>- 3.54KW (Cisco LAN switches)</td>
<td>- 6.1KW (DASD)</td>
</tr>
<tr>
<td>- Etc...</td>
<td>- 21 KW with disk</td>
</tr>
<tr>
<td>- 24.5 KW before adding disk</td>
<td>- Consumption fixed from 1 to thousands.</td>
</tr>
<tr>
<td>- Consumption increases as number of systems increases.</td>
<td></td>
</tr>
</tbody>
</table>

Consumption increases as number of systems increases.

Questions?

- Tell the commercial vendors you want to see Linux for System/390 applications!
- IBM is committed to Linux on the 390 and zSeries.
- Don't forget to tell your IBM rep you think VM is critical to the success of Linux on the S/390!

Contact Info

Linux-related stuff:

dboyes@sinenomine.net  
+1 703 783 0438
How Many Linux Systems Can Dance on a Single S/390?

Gratuitous Rah-Rah Slide

VM & Linux:

Let’s Rock Some Worlds!