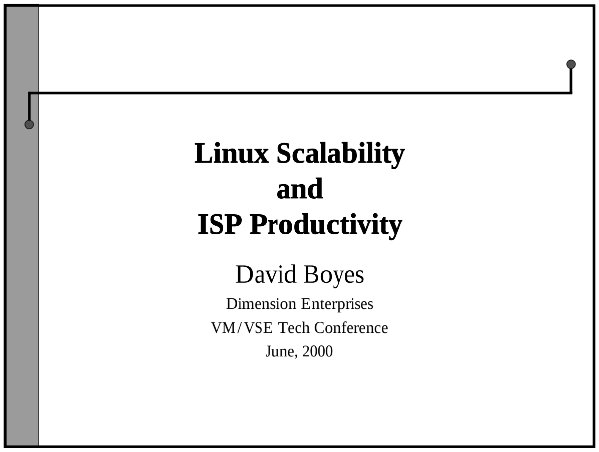
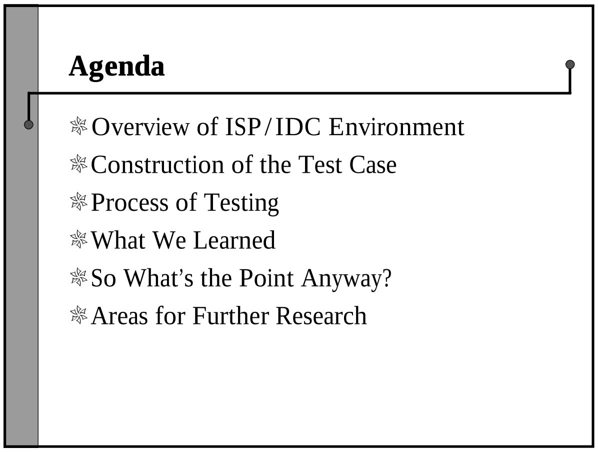


Linux Scalability and ISP Productivity

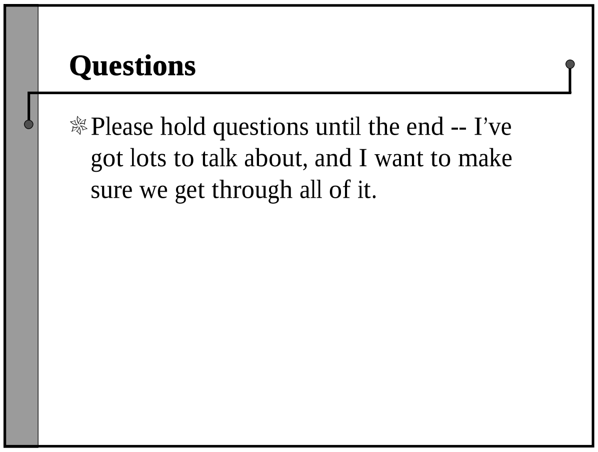
David Boyes
Dimension Enterprises

VM/VSE Tech Conference
June, 2000



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.

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, environmentals, operation and management software, etc.

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

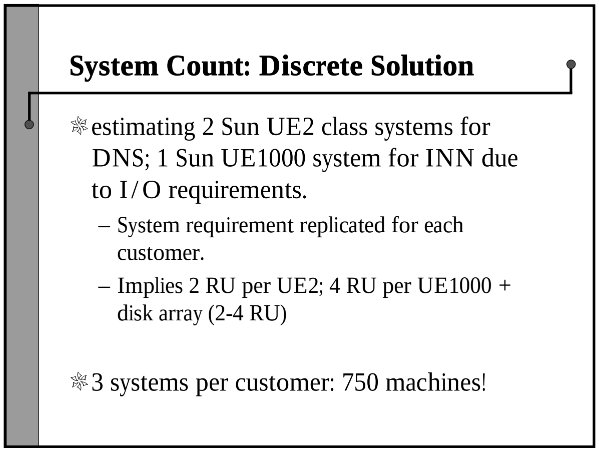
- Horizontal:
  - well suited to distributed apps and client/server
  - use of load balancing hardware hides complexity

Horizontal Vs Vertical Scaling

- Vertical:
  - well suited to interactive user sessions and applications
  - simpler to configure due to smaller number of machines

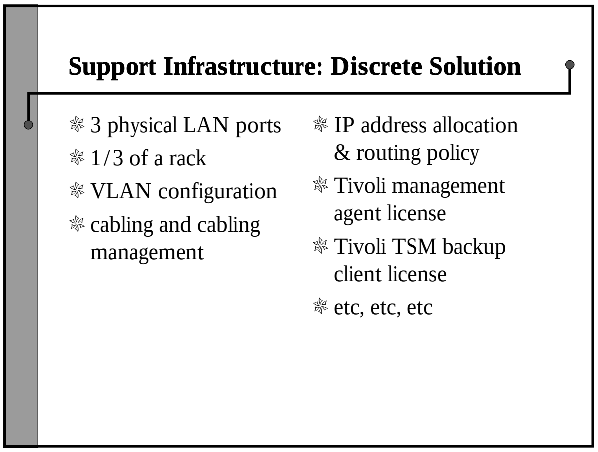
“Well, this is a pretty mess you’ve gotten us into…”

- 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)



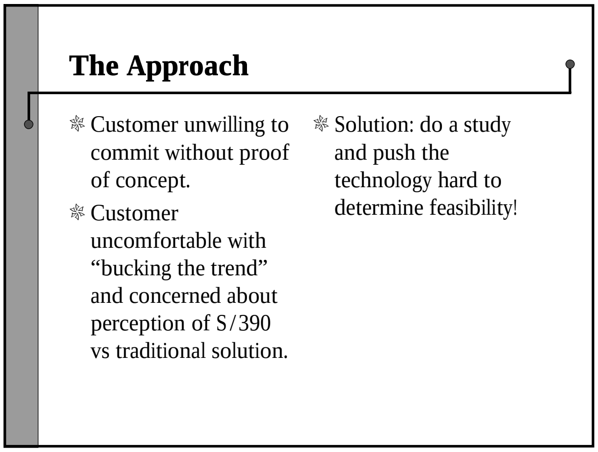
System Count: Discrete Solution

- 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!



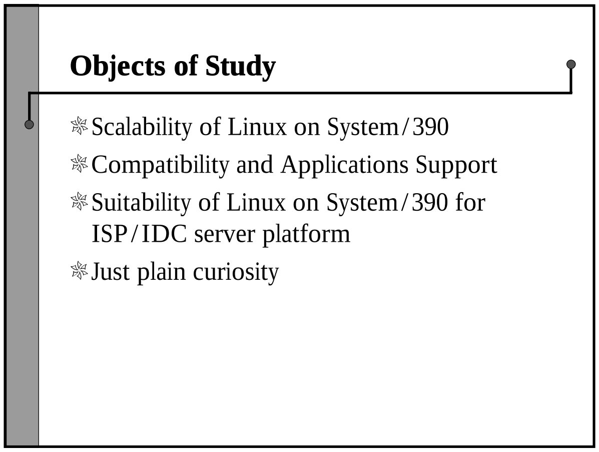
Support Infrastructure: Discrete Solution

- 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



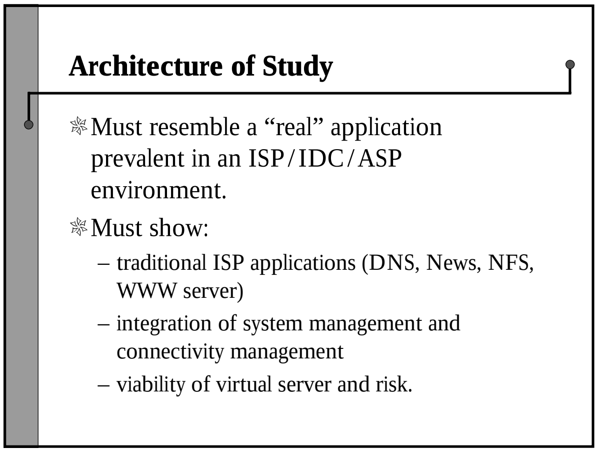
The Approach

- 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!



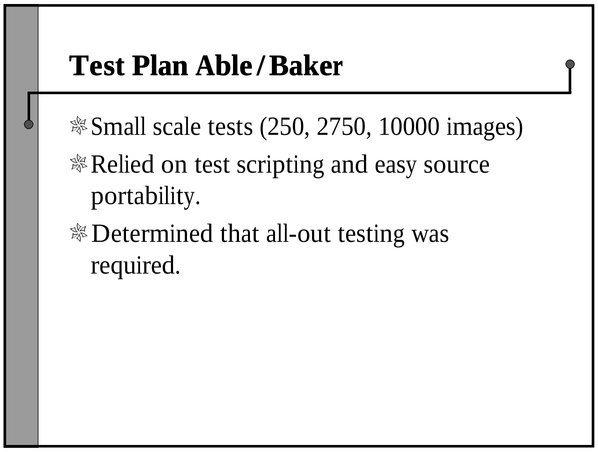
Objects of Study

- 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.

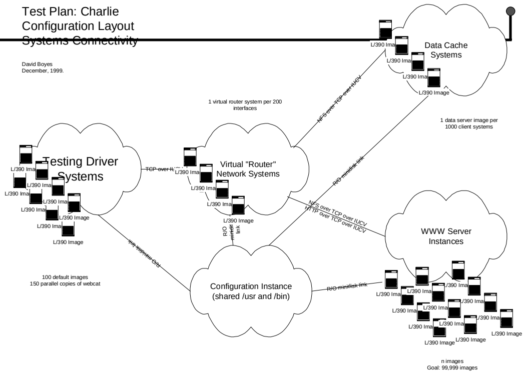


Test Plan Able/Baker

- 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
Configuration Layout
Systems Connectivity

David Boyes
December, 1999.

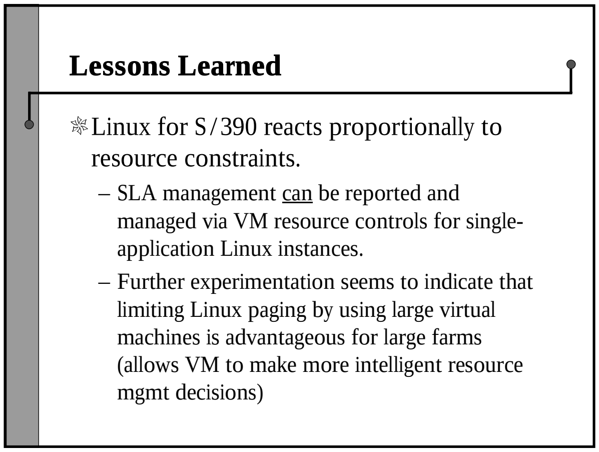


Lessons Learned

- 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
  - networking hardware management

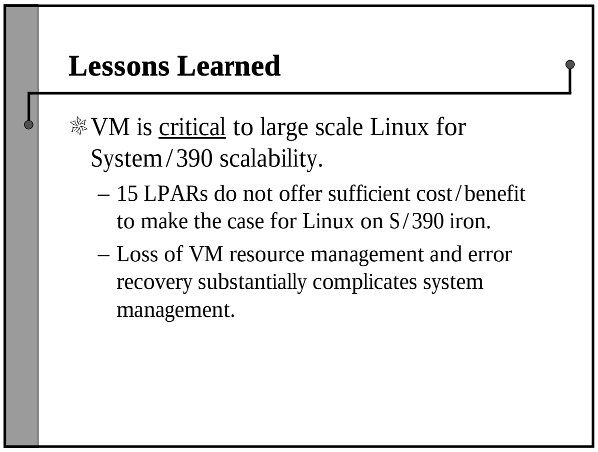
Lessons Learned

- 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.



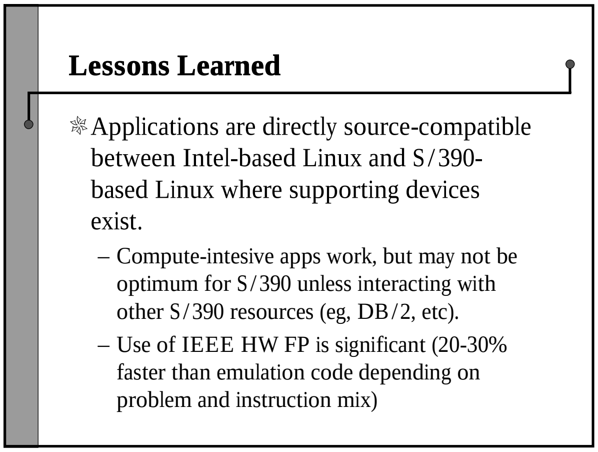
Lessons Learned

- 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 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.



Lessons Learned

- Applications are directly source-compatible between Intel-based Linux and S/390-based Linux where supporting devices exist.
  - Compute-intesive 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 almost completely automated (integrated into WWW front-end and back-end business systems).

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
- 1 high-capacity network cable (DS3/OC3/OC12 plus ESCON cabling to Cisco 7xxx+CIPs)
- 1 power cable per cabinet.

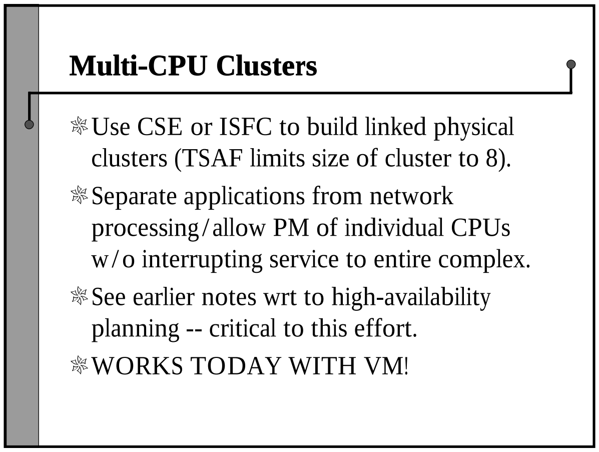
Simplicity!

Where to Go?

- Test Plan Omega: 100,000 images.
- Multi-physical box clustering
- Global clusters
- “VM Stun” -- migration of virtual machines between physical complexes.
- Non-S/390 Virtual Machines

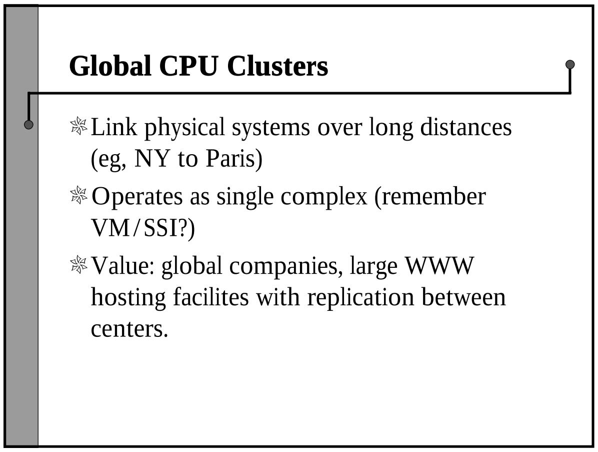
Test Plan Omega

- Push a single S/390 system to the limit: 100,000 systems
  - Endicott says that VM is supposed to support it as a design target -- let’s find out!
- Object: find out how many Linux systems we can cram onto one big box.
- Just looking for spare time to work on it. Anybody got a spare ZZ7 they’d like to lend some standalone time for this?



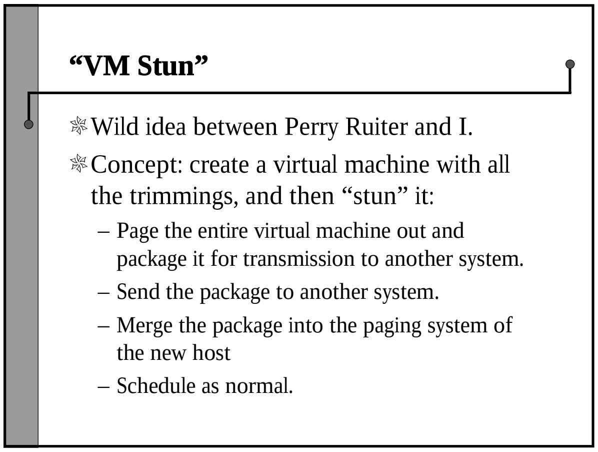
Multi-CPU Clusters

- 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 (eg, NY to Paris)
- Operates as single complex (remember VM/SSI?)
- Value: global companies, large WWW hosting facilites with replication between centers.



"VM Stun"

- Wild idea between Perry Ruiter and I.
- Concept: create a virtual machine with all the trimmings, and then "stun" it:
  - Page the entire virtual machine out and package it for transmission to another system.
  - Send the package to another system.
  - Merge the package into the paging system of the new host
  - Schedule as normal.

"VM Stun"

- Full suspend and resume capability without IPL of virtual environment.
  - Very, very difficult problems to solve here.
- Snapshot initiation of fully configured system w/o IPL startup configuration.
  - Very, very difficult problems to solve here too.

Non-S/390 Virtual Machines

- Why should VM emulate only the S/390 architecture?
- Can be done SLOWLY today with Linux for S/390 for almost any popular micro architecture:
  - Intel 486 (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.
- Future microcode bonus? X3?

Questions?

- ❄Don’t forget to tell your IBM rep that you want to see more Linux for S/390 apps!
- ❄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:

dboyes99@hotmail.com
+1 703 783 0438

Available in the Expo
somewhere near the Linux for S/390 booth.

Gratuitous Rah-Rah Slide

VM & Linux:

Let’s Rock Some Worlds!