CME Group

A CME/Chicago Board of Trade/NYMEX Company

CME Group Migration to Linux on x86: History and Challenges

Vinod Kutty, CME Group

2008/10/28 - IPA FORUM 2008

Photo: Wikimedia Commons,
Einar Einarsson Kvaran

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Intro – Overview

- Who we are
- Migration to Linux on x86
- How Linux/x86 has helped us
- Business Challenges
- Platform challenges and how we use Linux x86 Servers
- Future direction
- Summary

Who are we? History ...

Futures and options – currencies, energy, and even weather.

Chicago Butter and Egg Board founded in 1898. Evolved into Chicago Mercantile Exchange (CME) in 1919

History of innovation with financial instruments, including the first financial futures

CME Globex launched in 1992 (the first electronic futures trading platform)

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Who are we? History (cont.)

2007 - CME merged with Chicago Board of Trade (CBOT) – founded 1848, forming CME Group

2008 – CME Group acquired NYMEX (New York Mercantile Exchange – founded 1892)

Today CME Group offers futures and options based on interest rates, equity indexes, foreign exchange, energy, agricultural commodities, metals, and even weather and real estate.

Largest derivatives exchange in the world.

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Our volume

More than 2 billion contracts traded per year, valued at around US$1,200 trillion.

Most of this is electronically traded

Most of the electronic order entry and market data is handled by Linux on x86 servers (rack mounted 2-socket)

We also have proprietary systems.

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Business Challenges for CME Group

Capacity

Reliability

Availability

Performance

- **Requires constant R&D. Don't necessarily want bleeding edge.**

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Technical Challenges for CME Group

Performance! Milliseconds matter. Deterministic, low-latency scheduling.

Low latency UDP multicast

Low overhead, production ready performance tools (Systemtap?)

High-density servers (not necessarily blades)!

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Technical Challenges – x86 server hardware

Happy with hardware we have now, but need better standard *lowest-common denominator* hardware and associated tools designed from the ground up to run a CLI-based OS like Linux, not Windows!

- Also applies to other *nix platforms like Solaris x86
- more on this later

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Linux (RHEL) migration

Why?

- Compelling price/perf of x86 volume servers and Linux
- Reduction of support costs (hardware and OS)

From Solaris/SPARC – in 2003, starting with RHEL 2.1

Treated as just another Unix flavour. Manage + train personnel accordingly

Started with less critical non-customer facing apps

Then targeted more core apps

Ran some with primary and backup on Linux + Solaris

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Linux (RHEL) migration (cont.)

Most internal apps certified in about 6 months or less; others up to 18 months

Most systems migrated by late 2004

Now at thousands of systems and growing.

Pain points mostly with monitoring, admin script changes (ksh) and performance metrics/tools.

Our apps: Use of Java and horizontal scalability helped.

Close relationship with distribution vendor (Redhat) has enabled us to address most kernel issues over the years.

Training

- Not hard because most sysadmins already knew commercial Unix (Solaris) and were senior engineers
- some differences in tools needed adjustment
- about 6-8 months for about 15 people to be fully trained, but it was not hard to start administering Linux systems with little training
- did not trust Linux at first (2003). Acceptance grew as we saw price/performance benefits.

Reliability + Horizontal Scalability

- **Hardware features are important, but we rely more on application-layer fault tolerance**
  - allows greater customer control
- **Horizontal scalability (rather than vertical) is extremely important to us, and fits the commodity solutions model better**
- **Multi-core CPU trends favour horizontally scaled applications**

How Linux on x86 “volume servers” has helped

Price/performance much better than proprietary Unix/RISC platforms.

Initially, faster TCP/IP stack compared to STREAMs

Support is cheaper – hardware and operating system

More choice for support: OEMs or distro vendor

Our apps can scale horizontally; good fit with 1U server model

Good hardware RAID-1 w/ chassis swap support has eased maintenance

Price/performance benefits

- **Large systems with high-reliability at the hardware level are expensive long term investments**
  - Vendors invest significant engineering resources and pass costs on to us
  - latest CPU/Mem/Disk/IO not always available
- **1 or 2-socket x86 systems evolve rapidly and provide best value**
  - extremely competitive market segment
  - almost immediate access to latest technology every 18 months or so
- **Improved reliability at application layer helps decrease need for extreme hardware reliability, whilst benefiting from price/perf of 1 or 2-socket x86 servers**

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Other cost savings

- moving to a dual-vendor strategy for hardware
- competition between Linux distributions has helped lower support costs. Investigating free Linux alternatives for non-critical areas.
- more leverage for customers during negotiations, because we have a choice of vendors for hardware and operating system (compared to commercial Unix hardware + OS)
- simplified hardware maintenance + parts replacement (more later)

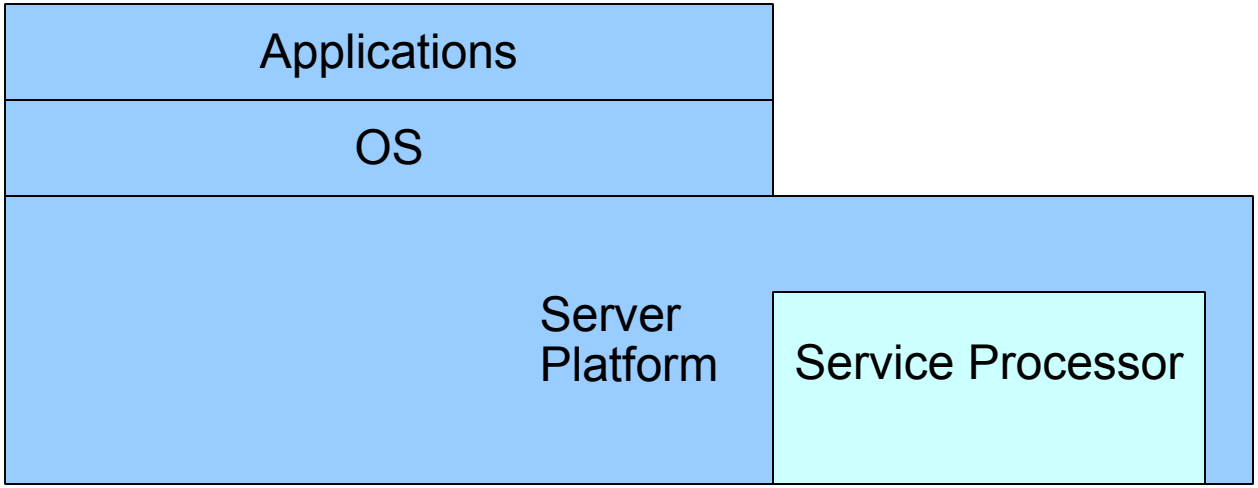
Support

- **Hardware parts replacement – self-supporting; keep spare parts in stock**
  - swap out entire chassis if there is a system problem, instead of waiting to diagnose issues

- **Linux –**
  - **1st: Internal self-support** (based on experience and online bug reports)
  - **2nd: work with our distribution vendor (Redhat) and hardware vendors**
  - for serious issues, **can look at source code** with our vendors, or ask vendors for existing upstream patches to be integrated

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Rough view of stack



Expect base platform to be headless (no framebuffer interface)

Fundamental Problems

Enterprise Ready Volume Servers are not true commodities.

CPUs, memory, I/O buses and some chipsets are, but little else.

Strong heritage of DOS/Windows

A desktop peg forced into a server hole

- Desktop heritage may lead to ignorance of “server” functionality and acceptance of weaknesses that can't be tolerated elsewhere

“Our servers fully support Linux” - what does this mean?

For many OEMs: “You can install and run Linux on our hardware.”

And it ends there. No substance behind it.

What about your hardware is designed to run Linux?

- Answer: nothing

“Happens to run” is not a substitute for “Designed to run”

Linux community works hard to support x86 hardware well

We have picked the cream of the crop in terms of offerings today, but the openness and Linux friendliness industry-wide needs improvement.

CME Group
A CME/Chicago Board of Trade/NYMEX Company

What does “Enterprise Linux Ready” mean to us?

General categories:

- Linux functionality
- Robustness of platform

More specifically:

- Linux == RHEL or SuSE, NOT latest kernel
- install out of box with no additional drivers
- full visibility into all aspects of hardware from Linux
- full serial console (physical and IP-based) support
- firmware updates from OS
- robust hardware design
- serviceability - good labeling, easy access to parts, etc.

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Hardware RAID is rocket science ... or so it would seem. What does CME depend on?

Many smaller OEMs don't understand RAID controllers

Some vendors claim hardware RAID, but aren't

Metadata is not stored on drives, so chassis swaps are not possible, or very difficult.

No optimizations for syncing drives that aren't full.

Poor tools for viewing/changing status/config from Linux CLI

We rely on ability to swap chassis and keep boot drives for simplified maintenance. Often no need to repair hardware – just replace.

CME Group
A CME/Chicago Board of Trade/NYMEX Company

(Hardware RAID continued ... examples of issues)

Cannot obtain event based notification via syslog of drive/volume status changes

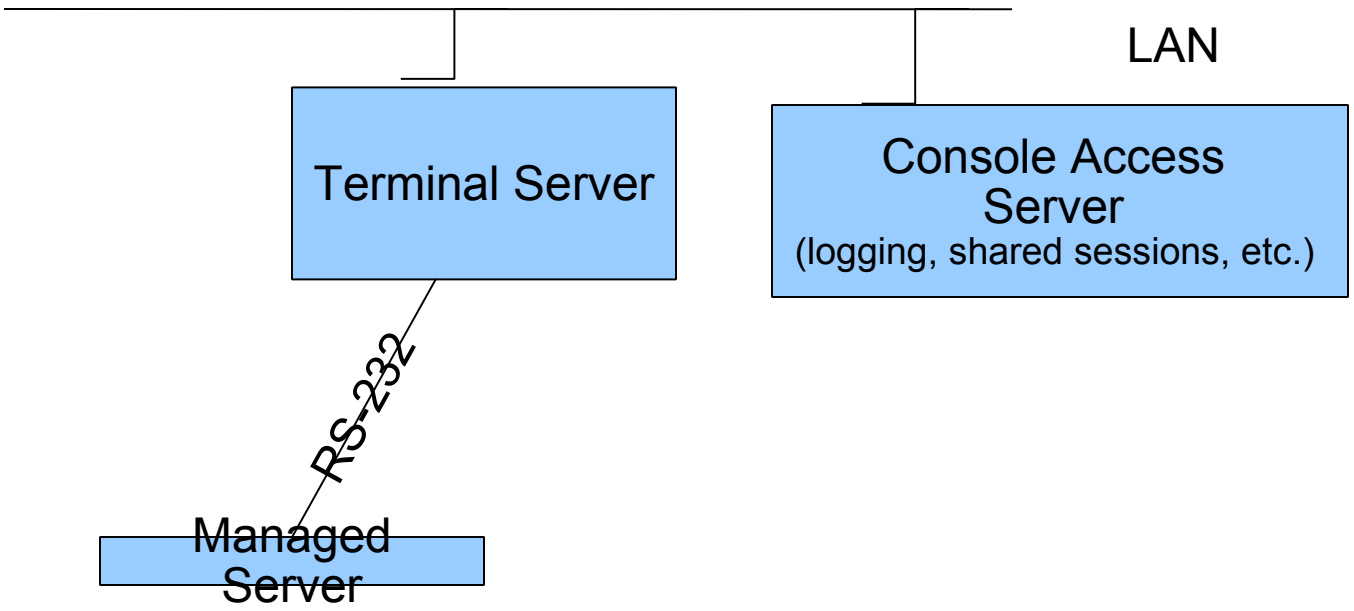
Cannot update RAID controller firmware from Linux

Cannot config RAID controller at POST via serial console (physical or IP-based connection).

Choose your vendor carefully!

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Serial Console (Physical)



Buffered console (history available)

Fewer network switch ports + IP ranges used

Terminal servers are robust – simple appliances; no moving parts

Uptimes measured in *years*

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Serial Console (Physical) cont.

Cost effective, simple, reliable – there when you need it. Well established in the commercial UNIX world.

Alternative: Framebuffers / KVM - a Windows-ism. Overkill, expensive to implement, no access to actual text output.

- imagine using screenshots of your desktop as a means of sending email

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Serial Console – IP based

IP-based mechanisms might be necessary for high density solutions, but are poorly implemented today (e.g. IPMI Serial-over-LAN)

- uptime certainly not measured in years
- Need a stable IP based solution before we'd move
- Need continuous connectivity to capture output (e.g. via *conserver* GPL solution)

Predictable PCI bus/slot device ordering

Example: Add quad-port PCI-e/PCI-x card to a slot. What happens to existing on-board NICs?

Depends on hardware and kernel

Frustrating

Hacks such as hard-coded MAC addresses have their own issues

Need something like the sbus-probe-list in OpenBOOT PROM

Newer pci=bfsort (kernel arg) may help

SMBIOS solution (proposed by Dell) may help

Need a standard guaranteed out-of-box solution

Firmware updates from within Linux

Should be able to update any firmware from OS:

- service processor
- BIOS
- RAID controller, etc.

Delivery as RPMs (Dell) is an interesting idea and one we like

BIOS

Why do we need it?

Would something else like Coreboot be more appropriate?

Need to be able to view/update settings from Linux using CLI

Need fast boot times and customized pre-boot environments

Replacing this is one innovation towards a more open, Linux-friendly platform

Hardware event monitoring

Need event notification that can be logged via syslog, for hardware status changes:

- failure/degradation
- recovery

Components must be clearly identified (e.g. 4GB DDR2 DIMM in slot 3)

Typically involves some daemon/agent that logs to syslog

IPMI

IPMI is a start, but has significant weaknesses.

IPMI model – BMC is a simple service processor, not designed for human interaction. So alternate service processors become necessary

KCS interface (for OS communication) is slow and seems to have not been a priority

Serial-over-LAN (SOL) not reliable

Peripherals optional so RAID subsystem events often not propagated

State of the x86 server industry from our perspective

A couple of large OEMs seem to have most but not all of our needs covered

- however, need proprietary stacks for best functionality
- moving to IPMI, but it lacks full feature set

Windows still commands largest market share, so OEMs wary of introducing features that may alienate them. Not a good way to win *nix market share!

Much of what we need has been around for at least 3-4 years or more

This is achievable!

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Future direction and Summary

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Firmware – coreboot + embedded Linux

- a truly open source replacement for proprietary BIOS
- bundled with embedded Linux
- fast boot times (3-5 seconds)
- familiarity (Linux) – re-use our existing tools
- more trustworthy for security reasons (due to open nature)
- easily customizable
- we are not interested in EFI/UEFI – why re-invent the wheel?
- will open up the x86 platform and make it more pervasive, following trend set by Linux

Overall better commodity hardware for Linux

- needs to be well designed for Linux with enterprise class features in low-end hardware
- learn from older commercial Unix systems
- better hardware event and management capabilities via CLI from Linux
- better RAID controllers
- better management processors with SSH and public key authentication

CME Group
A CME/Chicago Board of Trade/NYMEX Company

Better Linux system introspection tools for performance diagnostics and debugging

- some deficiencies today
- working with vendors and Linux community via Linux Foundation

More active Participants in Linux Community

- All users are guinea pigs for technology. We are essentially Quality Assurance (QA) for large vendors even though we may not realize it.
- Our time and money go to a small set of vendors, who keep the knowledge of our requirements to themselves
- As consumers, we replicate that effort and do not learn from each other, or benefit when we switch vendors.
- If we help improve Linux (and other open source solutions), that knowledge is not wasted, and will improve the state of the industry as a whole
- Almost like Wikipedia, with more moderation of input
- Also allows us greater involvement in our future

Summary

Linux on x86 has proven to be the most cost-effective platform for a wide variety of our applications

We are happy with the reliability and performance today

We continue to see cost savings due to increased competition from Open Source solutions such as Linux and amongst x86 vendors.

Participation in the Linux community by large enterprises is important for our collective future.

Application reliability and horizontal scalability are just as important as hardware and OS.

Need *lowest-common denominator* hardware and associated tools designed from the ground up to run a CLI-based OS like Linux, not Windows! coreboot + Linux are important to us (not EFI/UEFI)

Email: Vinod.Kutty at cmegroup.com

CME Group
A CME/Chicago Board of Trade/NYMEX Company