KVM on Embedded Power Architecture Platforms

Stuart Yoder
Software Architect, Freescale Semiconductor
Agenda

► Background
  • Freescale / Networking
  • Embedded Systems
  • Use Cases

► KVM on Embedded Power
  • New requirements
  • Status

► Future / To Do
Freescale: Networking & Multimedia Group

2010 Freescale Revenue

- Networking & Multimedia
- RF, Analog & Sensors
- Micro-controllers
- Cellular
- Other

NMG Revenue by Market

- Service Provider
- Enterprise/Small/Medium Business
- Industrial
- Multimedia
- Other

Key Networking Customers

- JUNiPER NETWORKS
- Alcatel-Lucent
- ADVANTECH
- Nokia Siemens Networks
- FUI Xerox
- HUAWEI
- EMERSON Network Power
- BROCADE
- CISCO
- SONY
- ZTE 中兴

Freescale is #1 in the network/communications processor market (300+million units shipped since 1989)
| QorIQ P5  | 64-bit High End  
Up to 2.2 GHz | Service Provider Network Admission  
Controls | Storage Networks | Switching |
|----------|-------------------|------------------|--------------|---------|
| QorIQ P4 | 4 – 8 Cores  
Up to 1.5 GHz | Metro Carrier  
Edge Router | IMS Controller | Radio Network  
Control | Serving Node  
Router |
| QorIQ P3 | 2 – 4 Cores  
Up to 1.5 GHz | Converged Media  
Gateway | SSL, IPSec,  
Firewall | Access  
Gateway |
| QorIQ P2 | 1 – 2 Cores  
Up to 1.2 GHz | Unified Threat  
Mgmt | VoIP Carrier-Class  
Media Gateway | Wireless Media  
Gateway | Base  
Station |
| QorIQ P1 | 1 – 2 Cores  
400 MHz to 1 GHz | Integrated  
Services Router | Network Attached  
Storage | Home Media  
Hub | Enterprise  
WAP |
Virtualization — Trends

- Desktop
- Enterprise/Datacenter: mainframes, servers
- Mobile
- Embedded: Aerospace, military (separation kernels)
How is embedded different?

- Fixed function devices— not general purpose
- Huge variety of hardware platforms
  - No standard platforms (no BIOS, ACPI, UEFI)
- Real time constraints
- Large variety of operating systems
  - VDC Research (2011 report)
    - About 50% of devices shipped by survey respondents had no formal OS or an in-house developed OS

Trend: move to multi-core SoCs, but SMP with a single OS will not be the only usage model
Trend: Consolidation on Multicore Processors

Benefit: Cost/power savings
Use Cases/Examples

► Control-plane / data-plane – split into partitions

► Migration — move to new hardware, preserve investment in software
  • Run legacy software alongside new software
  • Add Linux® to a system

► Sandbox — isolate untrusted software
Use Cases/Examples…continued

► High availability — active/standby configuration without additional hardware

► In-service upgrade

► Many more possibilities…
Standards

► power.org ePAPR
  • Resource discovery (device tree)
  • Multi-CPU boot
  • v1.1 includes virtualization extensions
    ▪ ABI
    ▪ APIs (hcalls)

► Power ISA 2.06B
  • Virtualized implementation notes
Why KVM for embedded Power Architecture?

Our customers are asking for it.
KVM on Power -- History

- **2007-2008:**
  - IBM developed 4xx processor (Book-III E) support (Hollis, Christian)

- **2009:**
  - Freescale did preliminary port to e500v2 (Yu Liu)

- **2009**
  - Port to server Book III S (Alex Graf)

- **2010-2011**
  - In progress: port to e500mc, improve/consolidate e500v2 work
New Mechanisms we need

► Assign guests physically contiguous memory
  • e500 MMU – software managed
    ▪ TLB0 – 4KB mappings
    ▪ TLB1 – small number of variable sized, large pages
  • Needed for performance (e.g. 80% speedup in kernel boot time)
  • Required for pass-through I/O devices to do DMA
    ▪ Freescale IOMMU supports a small number of DMA windows per device
    ▪ Devices with no IOMMU (e500v2-based)

► Pass-through of SoC I/O devices (non-PCI) to guests
KVM – e500v2

User mode
MSR[PR]=1

Kernel mode
MSR[PR]=0

QEMU

App

App

Guest

OS

Linux®
Kernel

MPIC

KVM

magic page
KVM – e500mc

- **User**
  - MSR[PR]=1
  - MSR[GS]=0

- **Hypervisor**
  - MSR[PR]=0
  - MSR[GS]=0

- **Guest Kernel**
  - MSR[PR]=0
  - MSR[GS]=1

- **Guest User**
  - MSR[PR]=1
  - MSR[GS]=1

- QEMU
- KVM
- Linux® Kernel
- MPIC
- Guest OS
- App

Freescale™ and the Freescale logo are trademarks of Freescale Semiconductor, Inc. All other product or service names are the property of their respective owners. © Freescale Semiconductor, Inc. 2009.
Status Summary

► Initial ports to e500v2 and e500mc based SoCs are complete
  • Basic features are there– sufficient to boot Linux® guest
  • e500v2 uses paravirt– shared page of memory and guest side patching
► Prototype direct map (pass-through) support for memory and I/O devices is working
  • Use in-kernel MPIC
► Upstreaming in progress
To Do

- Patches --> upstream
- Performance analysis & tuning
- Get rid of static guest device tree files
- Work out an improved mechanism to pass-through non-PCI I/O devices and physical memory
  - Hugetlbfs
- IOMMU support for SoCs with a PAMU
- Guest SMP
- 64-bit support (e5500)
- Additional VCPU features— e.g. debug, perfmon, cache locking
To Do…continued

- Error management
- Real time
- High availability
- Inter-partition communication/doorbells
- Direct hardware interrupts to guest OSes for pass-through devices
- Virtual time
- Libvirt
- Processor Roadmap
  - e6500 – has hardware threads and LRAT (logical to real address translation)
Conclusion

► Partitioning/virtualization is here to stay in the embedded space

► With some modest changes, KVM addresses many of the requirements

► Freescale sees direct customer demand for KVM and is committed to enabling this