Welcome

Connect Deutschland I.T. Symposium 2016

Moving to VSI OpenVMS V8.4-2 and beyond

Colin Butcher CEng FBCS CITP
Technical director, XDelta Limited
www.xdelta.co.uk
XDelta – who we are

- VSI Professional Services Alliance
- Independent consultants since 1996:
  - UK based with international reach
  - Over 30 years experience with OpenVMS
- We design and implement solutions:
  - Mission critical systems
  - Cross-sector experience
  - Engineering background
- Gartner (2009):
  - Identified XDelta as one of few companies world-wide capable of OpenVMS migration related projects
Personal background

• Systems architect specialising in mission critical systems

• Engineering background (printing presses, nuclear reactors, power generation)

• Wide range of experience (power generation and distribution, satellite control centres, air traffic monitoring, finance data, healthcare, transport, etc.)

• Started XDelta in 1996
XDelta – what we do

• Lead mission-critical systems projects

• Deliver world class services in demanding environments

• Strategic planning, technical leadership and project direction with clarity of vision and an eye for detail

• Systems engineering for availability and performance

• Ensure long term success through skills transfer
Agenda

• Why move to OpenVMS V8.4-2?
• OpenVMS V8.4-2 on Integrity Servers
• OpenVMS V8.4-2 on Alpha (evaluation kit)
• Infrastructure – hardware platforms, storage and networks
• Planning and design
• Examples and discussion
Part 1

• Why move to OpenVMS V8.4-2?
VSI OpenVMS – supported platforms

- x86-64 port in progress

- HP Integrity Servers:
  - Future “Kittson” family
  - Current hardware is -i4 “Poulson” family
  - Previous generation -i2 “Tukwila” family
  - Selected earlier server hardware (rx2660, etc.)

- HP Alpha:
  - “evaluation kit” recent release
# OpenVMS Rolling Roadmap

## March 2016

**OpenVMS V8.4-2**
- HPE Integrity System Support
  - Full BL890c Support
    - 64 Cores (hyperthreads off)
    - 1.5 TB Memory
  - AUTOGEN (large memory)
- UEFI 2.3
- Network Boot for i4 Blades
- WBEM for i4
- rx7640 / rx8640

**Performance**
- More alignment faults eliminated
- Tunable BACKUP Compression

**Software Component Updates**
- Enterprise Directory 5.7
- CSWS (based on Apache 2.4.12)
- LDDRIVER 9.7
- Latest Time Zone Definitions
- I18N (for localized language support)
- Digital Signing 2.0
- Rebranding 2.0

## 2016 / 2017

**OpenVMS V8.x,.....**
- HPE Integrity System Support
  - 16Gb Fibre Channel
  - HPE Kittson-based servers

**Software**
- JAVA 1.8
- VSI TCP/IP 1.0
- Larger Disks (64b LBNs)
- gSOAP, Samba, PHP, git
- OMNI / OSAP
- OpenVMS and LP Japanese kit

**New File System**
- In addition to ODS-2 and ODS-5
- Eliminate 2TB volume size limit, better performance and much more

**In Testing and Soon to be Released**
- 3PAR 8000/20000 Flash Arrays
- 64b Availability Manager
- OpenSSL 1.0.2g
- RTR

## 2018

**OpenVMS V9.0**
- Itanium & x86-64
  - Network stack performance
  - Containers
  - OpenJDK

**Itanium**
- Additional servers & I/O, depending on customer feedback

**x86-64**
- Selected HPE Servers (Intel and AMD)
- OpenVMS as a virtual machine guest
- Dynamic Binary Translator (Alpha/Itanium)
- Same compilers supported as on Itanium with standards updates for
  - C
  - C++
  - FORTRAN

---

Multiple releases may occur between V8.4-2 and V9.0. The order in which various improvements are added to these releases will be determined by readiness, hardware availability, and customer feedback.

These roadmaps contain forward looking statements and are provided solely for your convenience. While the information in this roadmap is based on our current best estimates, such information is subject to change without notice.
Migration to Integrity and beyond

Platform type / scale

Urgency

Small IA64 (eg: rx2660, rx2800-i2)  
Mid-range IA64 (eg: BL870c-i2)  
Large IA64 (eg: BL890c-i2)

Small Alpha (eg: DS25)  
Mid-range Alpha (eg: ES45)  
Large Alpha (eg: GS1280)

Small VAX (eg: microVAX)  
Mid-range VAX (eg: VAX 4000)  
Large VAX (eg: VAX 7000)  
HP VM
Part 2

- OpenVMS V8.4-2 on Integrity Servers
HP Integrity -i4 servers - highlights

• “Poulson” 2.53GHz 8 core processor with shared L3 cache

• Re-architected CPU design
  • Around 30% per core greater throughput
  • Reduced NUMA effects for same core count
  • Better memory latency and bandwidth
  • Improved floating point and integer performance
  • Improved hyperthreading – less stall time

• bl870c-i4 (32 cores) about 1.3x better than bl890c-i2
HP Integrity -i4 servers – hardware

- bl860c-i4:
  single width, 16 cores, 384GB, 4x 10GigE, 3x mezz, 1c2d SAS

- bl870c-i4:
  double width, 32 cores, 768GB, 8x 10GigE, 6x mezz, 2c4d SAS

- bl890c-i4:
  quad width, 64 cores, 1.5TB, 16x 10GigE, 12x mezz, 4c8d SAS
  - OpenVMS V8.4-1H1 currently supports a maximum of 32 cores
  - OpenVMS V8.4-1H1 also supports nPARs

- rx2800-i4:
  2U rack, 16 cores, 384GB, 4x 1GigE, 6x PCIe, 1c8d SAS
Server hardware differences (-i2 to –i4)

- Re-architected CPU, not just an updated design
- Higher clock rate (2.53GHz v 1.73GHz)
- “Out of order” instruction execution
- Minimal stall time between co-threads with hyperthreading
- Double the core count (8 cores)
- Greater memory capacity
- Reduced memory latency
- Shared on-chip cache

- New 10GigE LoM - LAN only, not FCoE
- Same 8GigFC mezzanine cards
Chassis hardware – c7000 / c3000

- Virtual Connect (GigE, 1/10GigE, 8GigFC)
- Flex10
- LAN side of FlexFabric

- 10GigE chassis based switching
- 10GigE passthrough
- 1GigE passthrough

- 8GigFC chassis switching
- 4GigFC passthrough
Infrastructure hardware

- 3PAR storage arrays at 8GigFC
- SSD devices for local storage and 3PAR storage arrays
- 8GigFC SAN – HP / Brocade switches
- 10GigE networking – HP Procurve, Cisco
VSI OpenVMS V8.4-2 on -i4 servers

- Complete build of base system from sources (V8.4-1H1)
- -i4 hardware support (64 cores supported, threads off)
- New LAN driver for LoM support
- VSI branding
- Patch kits available via HP
CPU architecture - Intel 9500 – “Poulson”
System architecture – rx2800-i4
Blade architecture – bl8x0c-i4
QPI fabric – bl870c-i4 and bl890c-i4
High core count

- CPU 00 is the primary CPU – try to reduce its workload
- Fastpath CPU selection – be aware of physical layout
- CPU choice for dedicated lock manager
- CPU choice for TCPIP packet processing engine
- Consider physical layout - RADs and NUMA
Hyperthreading

- Hyperthreading is extremely workload dependent
- In general the OpenVMS scheduler does a better job
- Enable / disable hyperthreads and reboot
- “CPU” count will appear to double when enabled
  - V8.4-2 will support 64 CPUs
  - V8.4-1H1 supports 32 CPUs (or “scheduling units”).
NUMA (non-uniform memory access)

- OpenVMS uses large shared memory regions:
  - XFC (50% available memory by default)
  - RMS global buffers
  - Global sections (especially database caches)
  - Memory disc driver (MD devices)

- Useful starting point for OpenVMS is “mostly UMA”
Memory architecture – bl890c-i4
Preliminary Performance Results: Integrity -i2 vs. -i4

- The following slides contain preliminary data on performance differences between selected i2 and i4 servers running OpenVMS E8.4-1H1.

- The data was generated from VSI-written programs used to measure certain aspects of system performance.
- The results shown here should not be used as a general characterization of overall system performance or as an indication of how any specific application may perform.
i2 vs. i4 Memory Bandwidth

- BL860c i2 (1.47ghz): 2968 MB/Sec
- BL860c i4 (2.4ghz): 5045 MB/Sec
- RX2800 i2 (1.73ghz): 2070 MB/Sec
- RX2800 i4 (2.67ghz): 4832 MB/Sec
i2 vs. i4 Memory Latency

![Bar Chart]

Latency (ns)

- BL860c i2 (1.47ghz) - 237 ns
- BL860c i4 (2.4ghz) - 196 ns
- RX2800 i2 (1.73ghz) - 243 ns
- RX2800 i4 (2.67ghz) - 108 ns
i2 vs. i4 Floating Point Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>i2 Performance</th>
<th>i4 Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL860c i2</td>
<td>1.00</td>
<td>1.33</td>
</tr>
<tr>
<td>BL860c i4</td>
<td>1.14</td>
<td>1.57</td>
</tr>
<tr>
<td>RX2800 i2</td>
<td>1.14</td>
<td></td>
</tr>
<tr>
<td>RX2800 i4</td>
<td>1.57</td>
<td></td>
</tr>
</tbody>
</table>

Blades: BL860c i2 (1.47ghz) vs. BL860c i4 (2.4ghz) vs. RX2800 i2 (1.73ghz) vs. RX2800 i4 (2.67ghz)
i2 vs. i4 Integer Performance

Relative Performance

<table>
<thead>
<tr>
<th>System</th>
<th>CPU Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL860c i2</td>
<td>1.47ghz</td>
</tr>
<tr>
<td>BL860c i4</td>
<td>2.4ghz</td>
</tr>
<tr>
<td>RX2800 i2</td>
<td>1.73ghz</td>
</tr>
<tr>
<td>RX2800 i4</td>
<td>2.67ghz</td>
</tr>
</tbody>
</table>
Performance engineering – use T4

• Avoid guesswork - run T4 all the time

• T4 “expert mode” and SDA extensions

• Without good data you cannot do good performance work

• A faster machine just waits more quickly
• Don’t make it go faster, stop it going slower
• The fastest IO is the IO you don’t do
• The fastest code is the code you don’t execute
Tips - OpenVMS V8.4-2 on -i4 servers

• Disable devices you don’t use
  SYSMAN IO SET EXCLUDE=(EWC,EWD,…)

• Experiment with memory interleave setting
• Use memory reservations

• Fastpath settings for device types
• Dedicated CPU for TCPIP + LCKMGR

• Experiment with hyperthreading
Migrating from Alpha to Integrity

- Multi-core processors, NUMA, hyperthreading
- 10GigE network
- 8GigFC SAN
- Blade chassis connectivity for bl8x0c-i4
- EVA to 3PAR storage migration
- bl870c-i4 and bl890c-i4: good for GS1280 migration
Benefits - OpenVMS V8.4-2 on –i4 servers

• Significant step up from AlphaServer GS1280, with modern storage and network infrastructure

• Significant step up from HP Integrity -i2 blade servers:
  o rx2800-i4 delivers more than 2x rx2800-i2 in same space
  o bl870c-i4 delivers more than bl890-i2 in half space
  o bl890c-i4 delivers more than 2x bl890-i2 in same space

• New VSI OpenVMS releases will bring new features and prepare the way for transition to OpenVMS on x86
Part 3

- OpenVMS V8.4-2 on Alpha (evaluation kit)
Now that we have released two new versions of VSI OpenVMS on the HPE Integrity hardware platform, we want to bring similar upgrades to Alpha hardware users,” said Duane P. Harris, CEO of VMS Software. “This evaluation kit will solicit feedback from Alpha users, and help us plan for a future Alpha production release.”

The Alpha Evaluation Kit is meant to reach a wide user audience. Their participation and input will be one of the key factors in deciding if VSI pursues OpenVMS releases for the Alpha platform. If VSI does proceed with future development, the Alpha release will follow the traditional Field Tests and release cycles.
VSI OpenVMS Alpha V8.4-2 Evaluation Kit

- Released April 2016, valid to September 2016
- Not for production use
- No support
- Base OS and network layers only - no layered products
- Only by upgrade from existing system

- Is there enough interest?
- How might it be put to use?
- What else would be needed?
- Please get in touch – we need your input
Potential ways forward with Alpha

- Extended support past HP deadline for organisations still reliant on Alpha, except VSI are not permitted to support non-VSI earlier versions (as with OpenVMS for Integrity)
- Supports both emulated hardware and physical hardware
- Emulated systems can be used in a virtual environment in limited circumstances
- Get current on Alpha before migrating to Integrity
- Buy time to complete migrations to Integrity
- Continue with Alpha until OpenVMS on x86-64 is available as a “production ready” system
- “Rights to new versions” is part of VSI licencing strategy
Some issues around staying with Alpha

• Becomes easy to put off moving to Integrity, which creates a bigger challenge later when moving to x86-64
• Emulators are becoming more capable, but big system performance will be hard to deliver from an emulator
• Hardware support (especially storage) becomes more difficult and is likely to be expensive
• Limited hardware capabilities of Alpha platforms:
  o 1Gbps or 2Gbps fibrechannel
  o 1Gbps ethernet
• Could new storage hardware be supported? Could it be supported as a boot device?
Recommendations

• If Alpha is of interest, get in touch and try the evaluation kit in a safe offline environment
• Migrating to Integrity now will help you migrate to x86-64 when it becomes “production ready”
• Understand the support issues and regulatory / legal implications if you choose to stay with old hardware
• Consider the surrounding infrastructure as well, especially storage subsystems
Part 4

• Infrastructure – hardware platforms, storage and networks
Scalable blades v rack mount

- Blades give expansion capability by adding blades (single to dual to quad). Can also be partitioned (NPARs).
- “Modular computing” - blades separate IO connections and power / cooling from machine

- Rack mount server is equivalent to single width blade
- No on-board 10GigE NICs
- Has PCIE card cage for expansion
- Internal SAS array with 8 drive slots in base system
Storage arrays - fibrechannel

• EVA now end-of-life
• HPE 3PAR, including all flash arrays
• HPE XP and XP7, including all flash arrays

• Arrays have different command interfaces, but broadly similar capabilities
• OpenVMS needs UUID value set when 3PAR virtual volume is exported (or EVA vdisk is presented)
• 3PAR now allows UUID to be changed

• SAN zoning connects HBAs to arrays
Network connectivity

- Multiple protocols: SCS, TCPIP, DECnet, AMDS
- Use LAN failover with multiple NICs for hardware resilience
- Use VLAN tagging and/or LAN failover sets to separate traffic flows
- VL / LL devices map to physical NICs, do not configure protocols on physical NICs.
- Use “service addresses” to separate data flows
- Use QoS in data network for different data flow types
- Use SCACP to control which port(s) SCS runs on
- Use LATCP to control which port(s) LAT runs on
- Disable unused protocols (eg: DECdns, DTSS)
OpenVMS networking: connectivity

- VLA (tag 1)
- VLB (tag 2)
- VLC (tag 3)

Switch A (tagged ports)
Switch B (tagged ports)
Part 5

• Planning and design
Design goals

- Design for change, not steady-state
- Operational safety – minimise risk of errors and disruption
- Understand the purpose and the target environment
- Build in logging and information gathering
- Adapt to changing requirements (performance, scalability)
- Think long-term (e.g.: company mergers / splits)
## Survivability – how long have you got?

<table>
<thead>
<tr>
<th>Cause of Outage</th>
<th>Planned (Maintenance)</th>
<th>Unplanned (Failure)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Operating System</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Network</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Application Software</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Data</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Environment</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>People</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Naming conventions

• Choose your naming conventions very carefully – they are the hardest thing to change later

• Don’t tie nodenames to physical locations

• Choose disc device IDs that identify meaningful things (e.g.: environment, site, array and purpose)

• Choose network addresses and hostnames that identify meaningful things and make sense in your context
Shadowing

- Many shadow sets for performance with multi-path discs
- Small shadow sets to minimise copy/merge time (especially common disc)
- Enough arrays per site to always have local source
- Only mount system discs on nodes booted from that disc
- System disc at a site is shadowed to other sites
- Use minicopy and minimerge for performance
- Consider having temporary extra members during changes
Quorum and voting

- Is application “cluster aware” or rapid failover?
- What do you want to happen when a site fails?
- Avoid quorum disc if possible
- Quorum adjustment and changes to voting scheme while nodes/sites are temporarily removed
Hardware maintenance and replacement

- Keep firmware up to date – plan sequence to avoid disruption

- FC devices with same UUID but different WWPNs will show up as the SAME device but with extra paths.

- Keep systems modular with minimal configuration per node

- Save / restore ILO configurations with USB flash drive
Availability manager

- Windows management stations (typically one per site)
- Gives “real time” view of nodes in management group(s)
- Uses AMDS protocol (layer 2 – use LL or VL device)
- Interacts with OpenVMS driver at high IPL
- Permits modification of running system:
  - Quotas
  - Dynamic parameters
  - Quorum
Planning the changes

- Permissible downtime usually controls the planning
- Risk of disruption usually controls the sequencing
- Never do too many things at once
- Only do one thing at once for critical actions
- Identify any one-way steps as early as possible
- Stay current enough to be supported
Getting things done safely

- Shift workload around to deliver continuous service
- Write everything down in painful detail as a checklist, with anticipated timings
- Rehearse as much as you can
- Do as much as you can in advance
- Make sure everything is tested and ready to go
- Tell everyone what’s going on and monitor everything
- Never proceed until you know things are safe
- Don’t be afraid to call it off and back out
- Never work alone
Techniques for rolling upgrades

- Primary and alternate system discs per site (or per node)
- Page / swap / dump files off system discs
- Common disc(s) for system-wide files etc. (UAF, rightslist, queue manager databases, etc.)
- Keep everything tidy and know where everything is
- Make copies in the storage arrays before you start
- Know how to back out and revert to where you started from
Part 6 – this needs your input

- Getting to V8.4-2: examples and discussion
- These are based on real-life scenarios
Starting from Integrity (1)

- A single production node (rx2660), non-clustered with an alternate system at another location
- Local SAS storage within the machine, Partitioned RAID6 p400 RAID controller (not the embedded controller)
- SAS tape backup
- Data replication by copying data over network
- Evenings / weekends available with planning ahead
Starting from Integrity (2)

- A multi-site cluster, two or more nodes per site (original blades: bl860c, not –i2 servers)
- Fibrechannel SAN storage (EVA)
- Boot from SAN with single system disc per site
- Local SCSI storage for temporary data only, RAID controller
- Fibrechannel tape backup using array capabilities for snaps & clones to minimise backup window
- Data replication by fibrechannel SAN extension with MSCP fallback
Starting from Integrity (3)

- A multi-site cluster, two or more nodes per site (32 way bl890c-i2)
- Fibrechannel SAN storage (mixed EVA and 3PAR)
- Boot from SAN with single system disc per site
- Local SAS storage for temporary data only, RAID controller
- Fibrechannel tape backup using array capabilities for snaps & clones to minimise backup window
- Data replication by fibrechannel SAN extension over DWDM with MSCP fallback
- Downtime very difficult to get. Need more performance.
Starting from Alpha (1)

- A single production node (ES47), non-clustered with an alternate system at another location
- Local SCSI storage within the machine, RAID controller
- SCSI tape backup
- Data replication by copying data over network
- Evenings / weekends available with planning ahead
Starting from Alpha (2)

- A multi-site cluster, two or more nodes per site (ES47)
- 2Gig Fibrechannel SAN storage (EVA)
- Boot from SAN with single system disc per site
- Local SCSI storage for temporary data only, RAID controller
- Fibrechannel tape backup using array capabilities for snaps & clones to minimise backup window
- Data replication by fibrechannel SAN extension using FC over IP with MSCP fallback
- Downtime possible with careful planning ahead
Starting from Alpha (3)

- A multi-site cluster, two or more nodes per site (32way GS1280, Galaxy), with an alternate cluster at another location for extra resilience
- 2Gig Fibrechannel SAN storage (EVA)
- Boot from SAN with single system disc per site
- Local SCSI storage for temporary data only, RAID controller
- Fibrechannel tape backup using array capabilities for snaps & clones to minimise backup window
- Data replication by fibrechannel SAN extension over DWDM with MSCP fallback, plus array based replication
- Downtime very difficult to get. Performance critical.
Summary

• Can we create a checklist of things to look for?
• Can we set up a standardised approach?
• Get current on old platform (equivalent versions to new)
• Clean up everything as much as possible on old platform before moving to new platform
• Data migration for historic data and backups / archives
Wish list!

• Single node cluster licence as part of base OS
• Single member shadow sets as part of base OS
• ALLOCLASS per storage array / tape library (WWNN based?)
• Do not start SCS / DECdns etc. by default on all NICs
• Single instance Galaxy memory on Integrity (and x86-64)
• What else? Let us know!
Thank you

Connect Deutschland I.T. Symposium 2016

Moving to VSI OpenVMS V8.4-2 and beyond

Colin Butcher CEng FBCS CITP
Technical director, XDelta Limited
www.xdelta.co.uk