Hewlett Packard Enterprise

# Hybrid workloads with NonStop

Prashanth Kamath U (HPE Product Management)
Thomas Burg (comForte 21 Gmbh)

April 18, 2016

# Forward-looking statements

This is a rolling (up to three year) Roadmap and is subject to change without notice.

This document contains forward looking statements regarding future operations, product development, product capabilities and availability dates. This information is subject to substantial uncertainties and is subject to change at any time without prior notification. Statements contained in this document concerning these matters only reflect Hewlett Packard Enterprise's predictions and / or expectations as of the date of this document and actual results and future plans of Hewlett Packard Enterprise may differ significantly as a result of, among other things, changes in product strategy resulting from technological, internal corporate, market and other changes. This is not a commitment to deliver any material, code or functionality and should not be relied upon in making purchasing decisions.

# HPE confidential information

This is a rolling (up to three year) roadmap and is subject to change without notice.

This Roadmap contains HPE Confidential Information.

If you have a valid Confidential Disclosure Agreement with HPE, disclosure of the Roadmap is subject to that CDA. If not, it is subject to the following terms: for a period of 3 years after the date of disclosure, you may use the Roadmap solely for the purpose of evaluating purchase decisions from HPE and use a reasonable standard of care to prevent disclosures. You will not disclose the contents of the Roadmap to any third party unless it becomes publically known, rightfully received by you from a third party without duty of confidentiality, or disclosed with HPE’s prior written approval.

# Agenda

– What was announced in GTUG 2015 – a recap

– NonStop Application Direct Interface (a.k.a. YUMA)

– NSADI possibilities – comForte

– Round Table

– Wrap up and the next steps

# IT Transformation

## Enterprise imperatives

- Speed innovation
- Accelerate services
- Improve flexibility
- Do more with less
- Manage risk

## Mega trends

- Big Data
- Cloud
- Mobility
- Security

Hewlett Packard Enterprise

# NonStop and Linux — a hybrid approach for the new style of IT

Tighter integration of classic and new environments

## Best of both worlds

**Hybrid Linux and NonStop environments have already been deployed**

**NonStop is making significant investments to enable a more seamless hybrid environment**

**Rock solid scalability**

**Availability and disaster recovery**

**New open source frameworks and features from Linux**

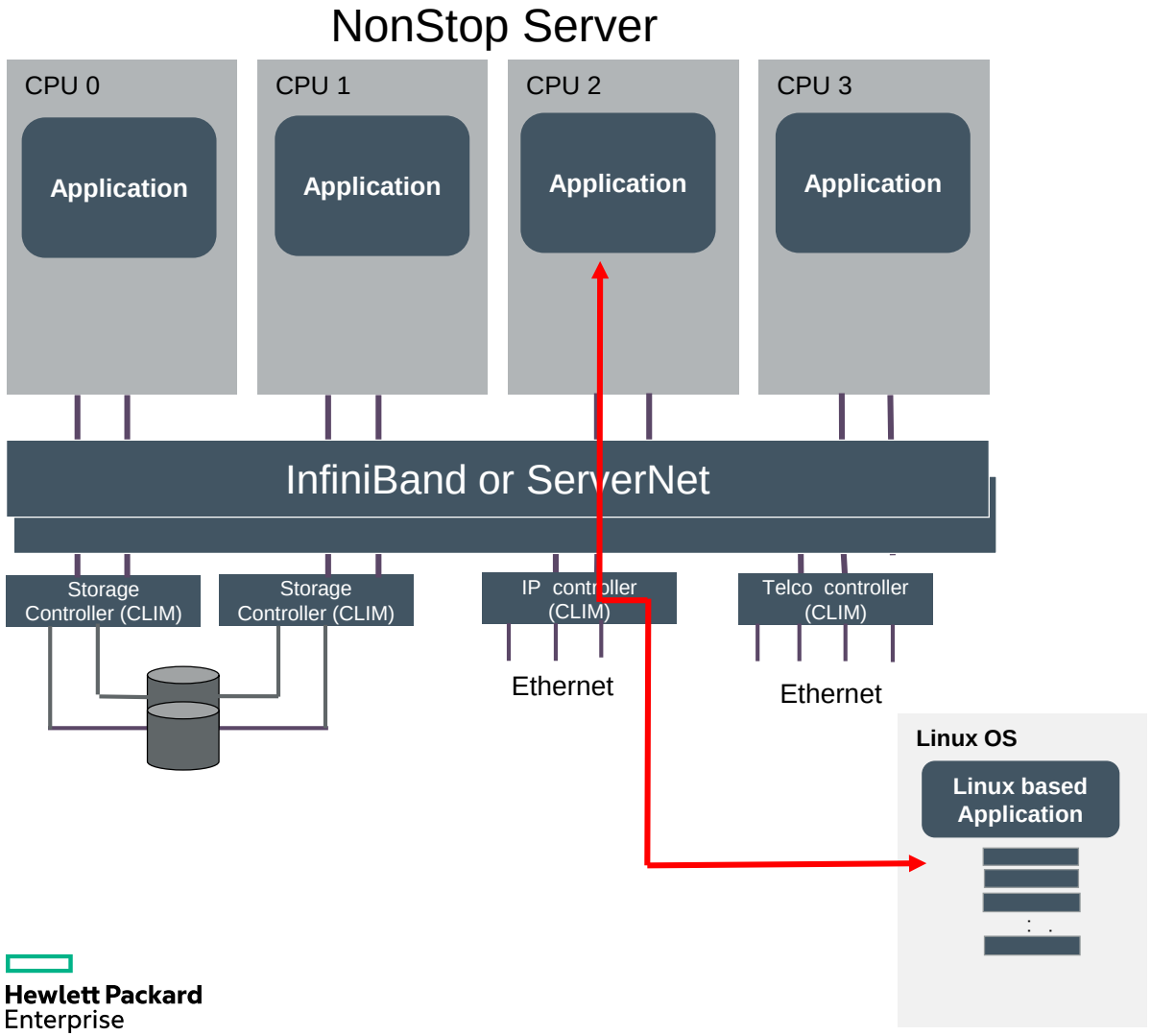
Hewlett Packard Enterprise

# Investing Beyond 2015 for the Virtualized Future

- NonStop has always been integrated in hybrid environments
  - Countless customer use cases and examples
- NonStop X provides more than a platform refresh to a new technology
  - Introduces InfiniBand, an industry standard – high bandwidth, low-latency interconnect
- InfiniBand allows creation of  seamless environments ranging across
  - Front-End / Back-end Hybrid environments
  - Private and Hybrid Clouds
  - Internet of Things
- New investment areas:
  - Hybrid
  - Virtualized Environments

# NonStop to Linux connectivity - Today

## TCP/IP Interface




This is a classic node to node connectivity over a TCP/IP network

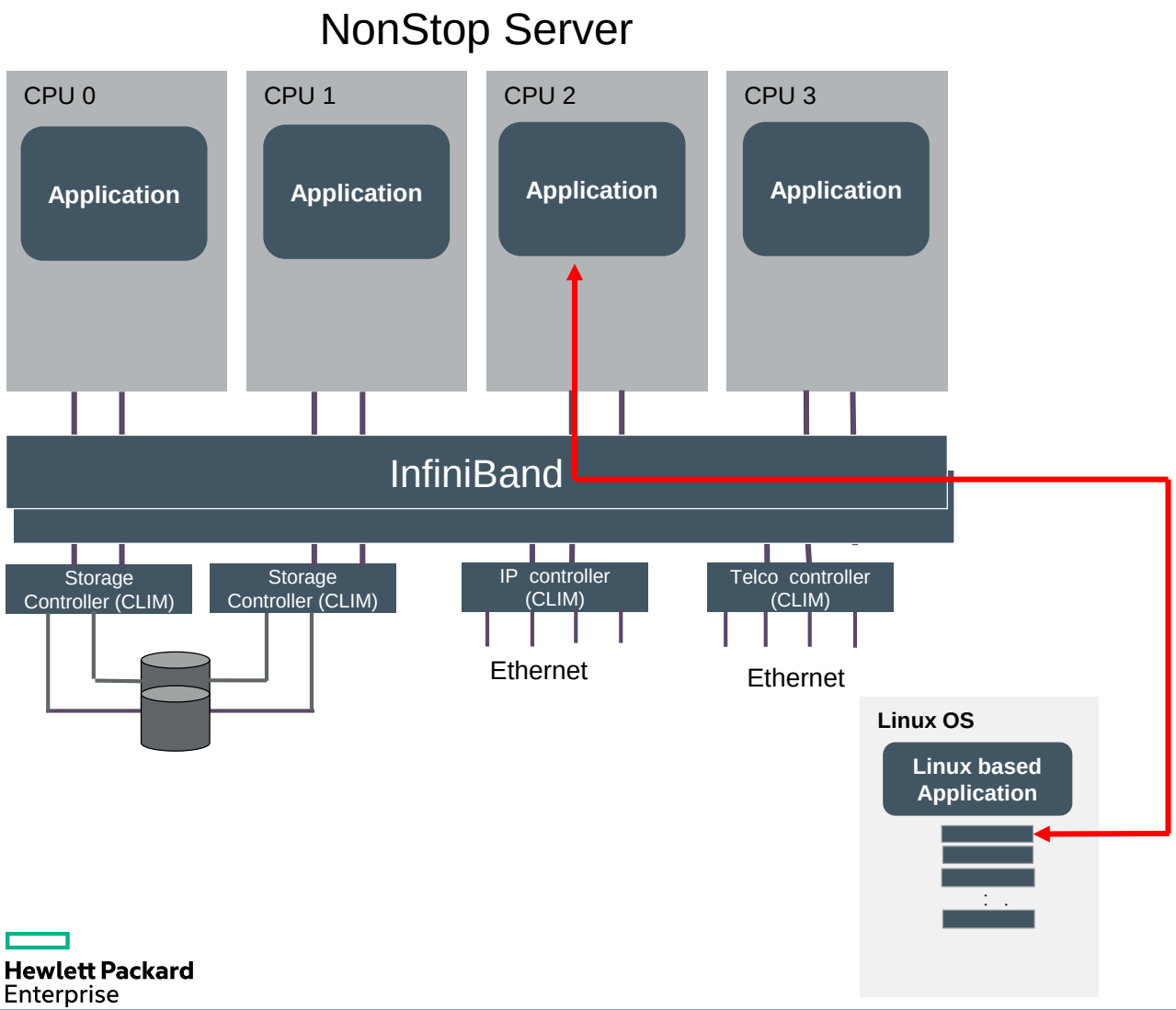
Involves multiple data copies and transport via a (slower) Ethernet link

Not suitable for solutions which need,

- Lowest possible latency
- Bulk data transfer
- Low CPU usage (on NonStop)

# NonStop to Linux connectivity - Future

## NonStop Application Direct Interface (NSADI)




Applications write to the user memory on the remote host using Remote Direct Memory Access (RDMA)

No copies between user and kernel buffers

Benefits

- Lower latency
- Better throughput
- Minimal NonStop and Linux CPU usage

Hewlett Packard Enterprise

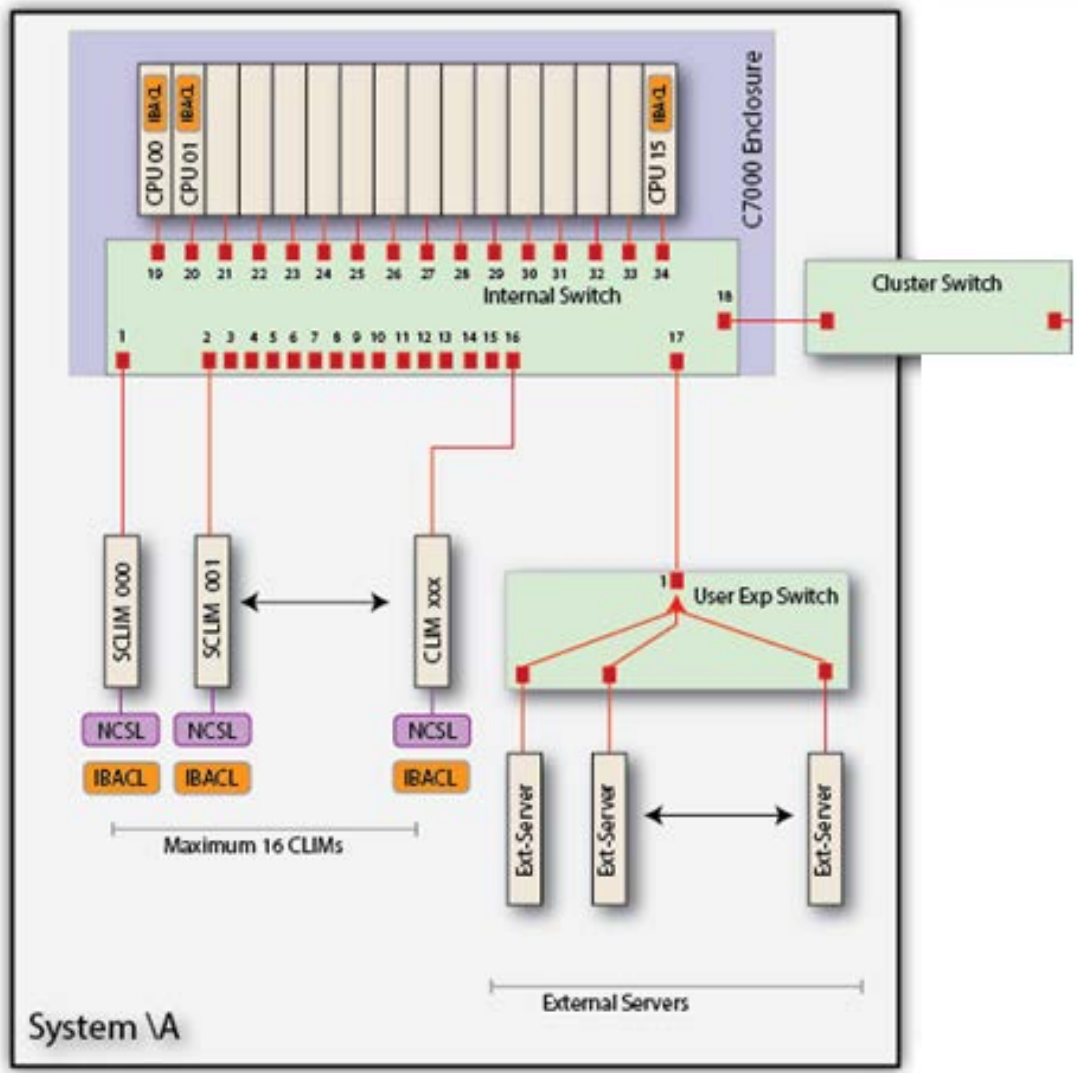
# NSADI

## User/Kernel mode interactions

- Designed to minimize application-to-kernel interactions
  - Data transfers do not require a privileged transition into the kernel
  - Data transferred directly into/out of user buffers. Does not require the kernel to copy data across the user/kernel divide.
  - Interrupts for received buffer indications (which require kernel interactions) can be minimized.
  - The kernel path for interrupt reception is very short. The kernel need only notify the user application that data is present in its buffers.
- Initial connection start-up and tear down do require kernel interactions.

# High Level Architecture

## Physical Connections




External Servers will connect to the NonStop system via a dedicated IB switch for NSADI connectivity

NonStop supplied processes labeled “IBACL” provide security by preventing the external servers from accessing critical data or subsystems on the NonStop CPUs or CLIMs.

Maximum NonStop CLIMs not affected by NSADI.

# High Level Architecture

## Overview

–NSADI bypasses the networking CLIMs for data exchange

- Allows a direct connection between the external servers and the NonStop user application.
- Data can be placed directly into the user memory buffers.
- No kernel interactions are required for bulk data IO

– Other parts of NonStop CPU memory and the CLIM based subsystems cannot be accessed from the external servers.

- Applications on the external servers will NOT be able to access the storage subsystem (customer data disks) or CPU memory on the storage CLIMs.
- Applications on the external servers will NOT be able to access the networking CLIMs via NSADI
  - Can access NonStop networking CLIMs via TCP/IP

# High Level Architecture

General InfiniBand Characteristics

- Highest levels of data Integrity
  - Cyclic redundancy checks (CRCs) at each fabric hop and end to end across the fabric
- High Bandwidth / Low Latency
  - InfiniBand provides increased bandwidth and low latency required for demanding IO centric applications on the x86 platform.
- RDMA
  - The ability to remote DMA data into/out of CPU memory without kernel intervention enhances efficiency of customer workload processing

# What’s coming in the first release?...

## Hardware

- NonStop to Linux (RHEL) connectivity over a dedicated IB switch
- Supported on High End (NS7) and Entry Class (NS3) systems
- Connect up to 8 Linux servers on NS7 and up to 2 Linux servers on NS3

# What’s coming in the first release?

## Software

NonStop applications using this architecture must be:

- OSS based applications
- 64 bit / PUT model

Application Programming Interfaces

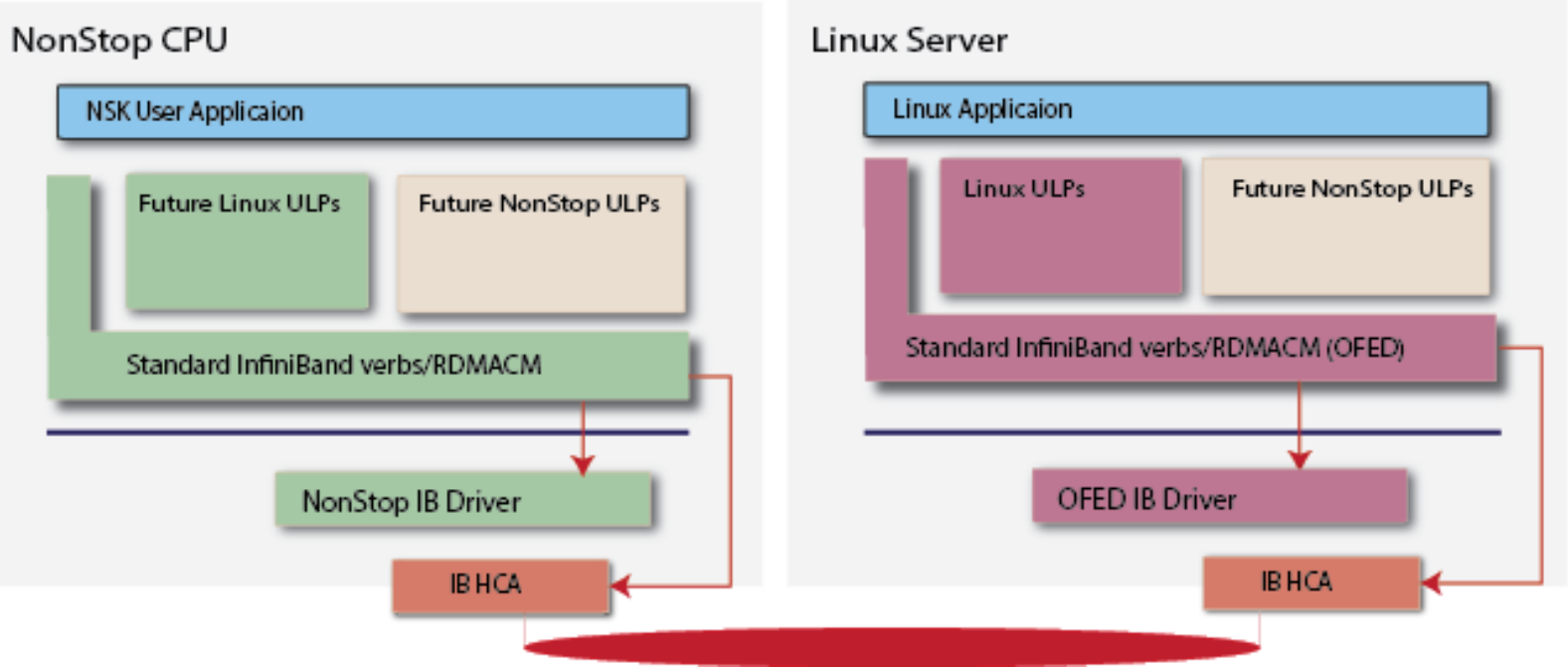
- IB Verbs: Lowest interface layer. Best throughput and latency; connection establishment and management done by the application
- RDMACM: Socket like interface adapted for queue pair based semantics. Used for connection management.
- RDMA Sockets (rsockets): Socket based interface. Aids portability; lower throughput and latency compared to IB verbs. Not much impact for large messages

The matching verbs/RDMACM Linux side components are open source libraries that are readily available on RHEL distributions (no cost).

Licensing: Optional, separately licensed product enabled through core license file

# Software Architecture

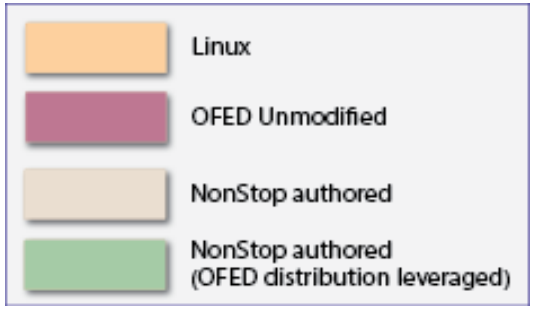
## Matching Stacks (Futures)



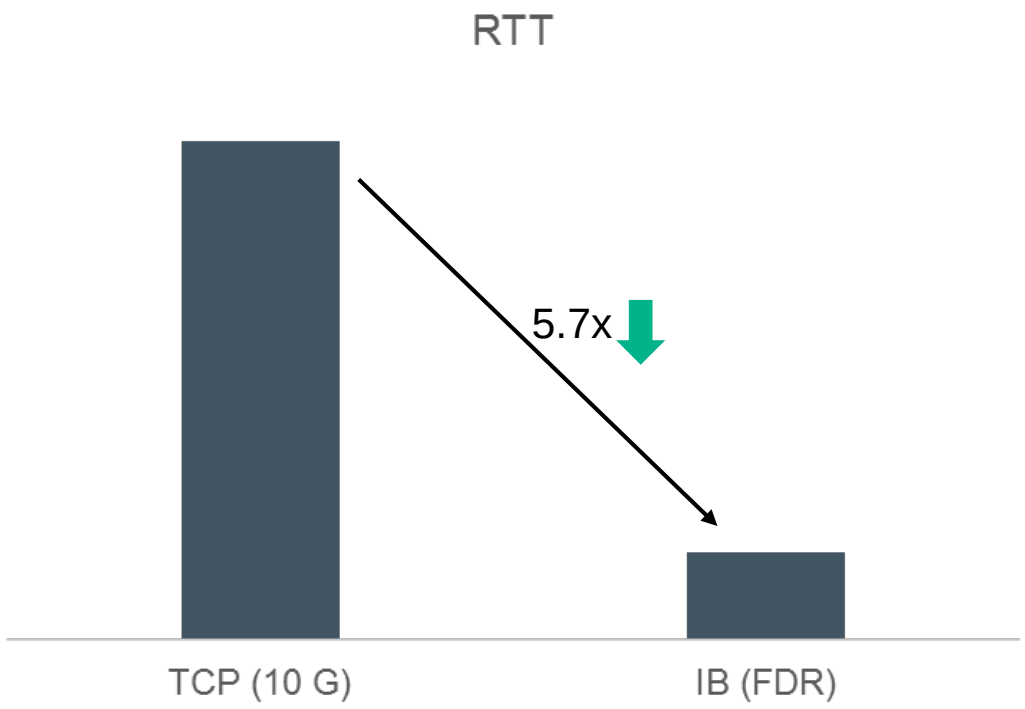

NonStop user mode InfiniBand provides matching layers to Linux servers.

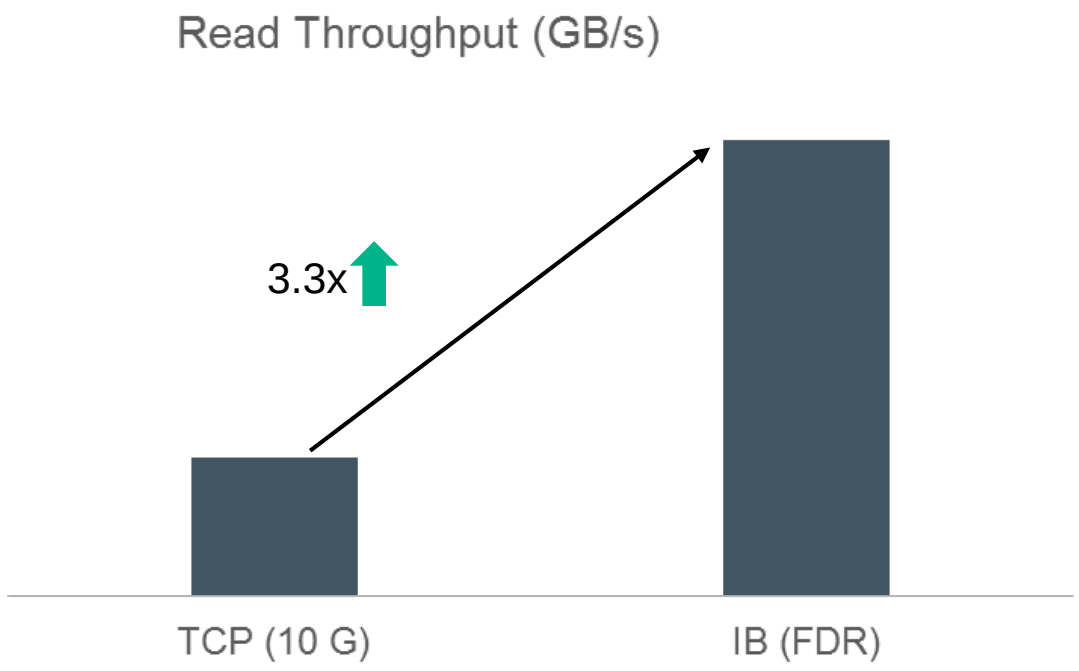
The user level verbs/RDMACM and Rsockets layers on Linux are the standard OFED distribution.

Requires no modifications.



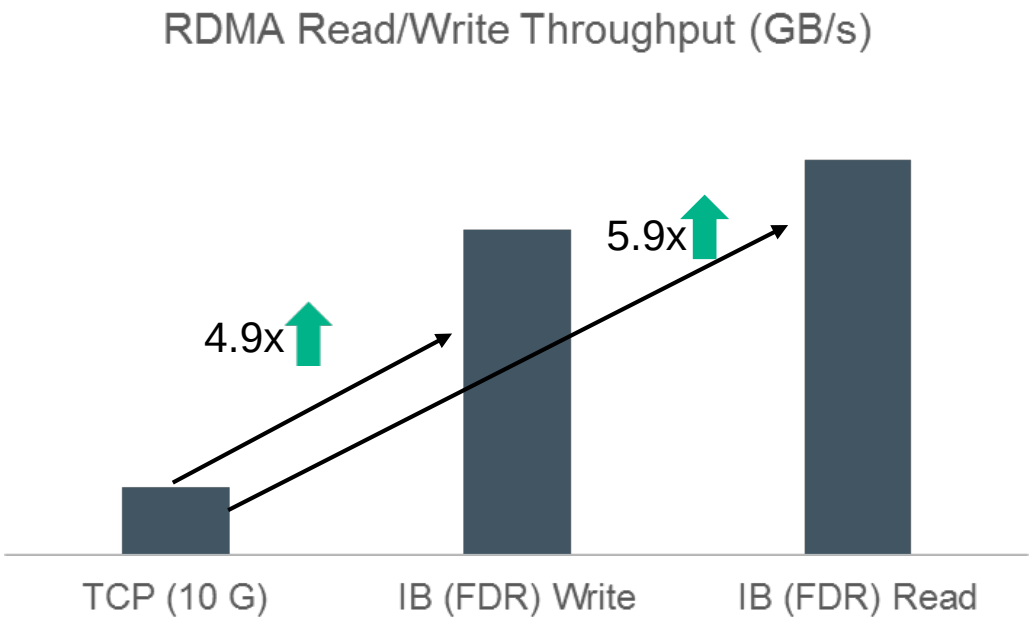
# (Very) Preliminary performance results* …

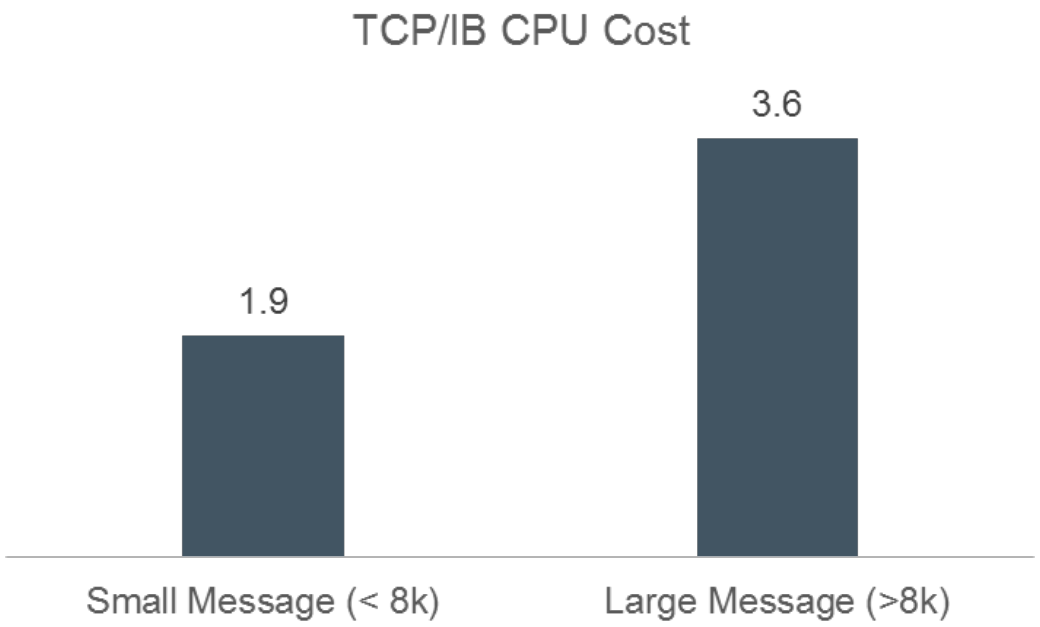



* Your mileage is expected to vary from these results

# (Very) Preliminary performance results*



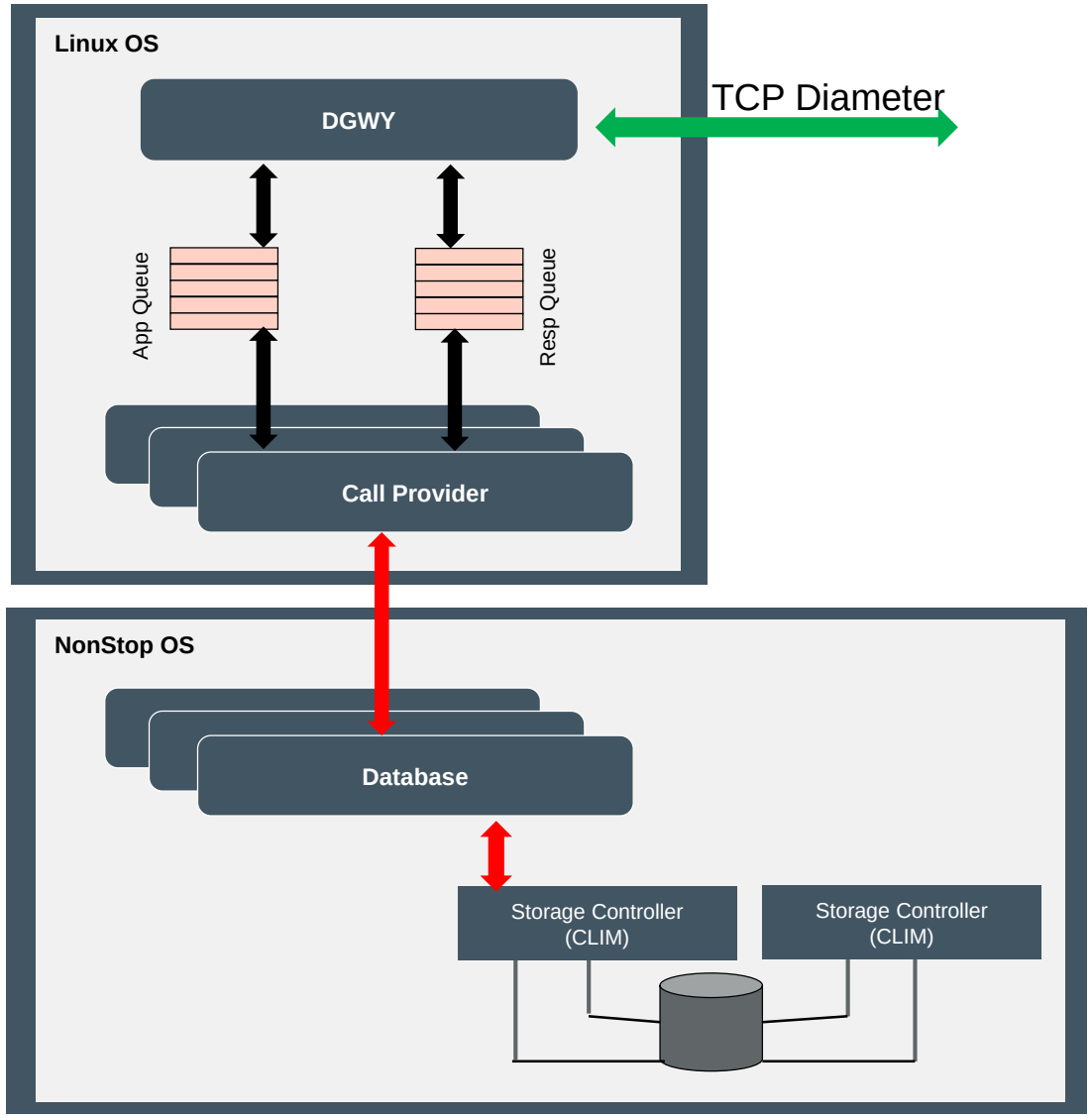




* Your mileage is expected to vary from these results

# HPE Integrated Home Subscriber Server (I-HSS)

## Proof Of Concept (POC)

- Incoming Diameter and ECPY messages are distributed by the DGWY process to one of N Call Provider processes.
- The call provider process interprets the Diameter messages and creates a LDAP transaction to its matching NonStop server process (database end-point).
- The existing Linux application was designed with a C++ class that performed all IO to the NonStop OS. This class was replaced with logic that performed InfiniBand verbs based IO (effectively hiding the transport from the overall application).
- Application round-trip time for a given message using NSADI is ~3.5x faster than the original TCP/IP based transport




Hewlett Packard Enterprise

Hewlett Packard Enterprise

# NSADI

## comForte proof of concept

# Yuma POC “phase 2” @ comForte

## Results
## Product visions

*Thomas Burg, March 2016*

better always on

com.forte®

# Agenda

- What is InfiniBand and why should you care
- The comForte Yuma POC phase 2 results
- A business case and technical vision for „Hybrid NonStop“
- The comForte Yuma product vision

better always on

com.forte®

# Typical speeds for TCP/IP networking (*)

"full roundtrip time"
= 0.4 milli seconds

= 400 micro seconds

= 2500 TPS

+

165 min to move a TeraByte

=

Current speed NSX over 1 Gb Ethernet

# Moving to InfiniBand, comparing to 1 Gbit TCP IP over Ethernet…

"full roundtrip time"
= 11 micro seconds

= 90.000 TPS

= 34 x faster

+

3 min to move
a TeraByte

= 55 times
faster

=

Awesome
Speed of IB

**(*) How fast is fast enough? Comparing Apples to ….**

# When do I need to move data really really fast …

BigData (Duh)

Stock Exchanges

Telco

NonStop Hybrid

Discussion to follow

# The comForte Yuma (a.k.a. NSADI)

## POC Phase 2

- Context
- Goals
- Results
  - **Moving data**
  - **Moving files**
- Other observations

better always on

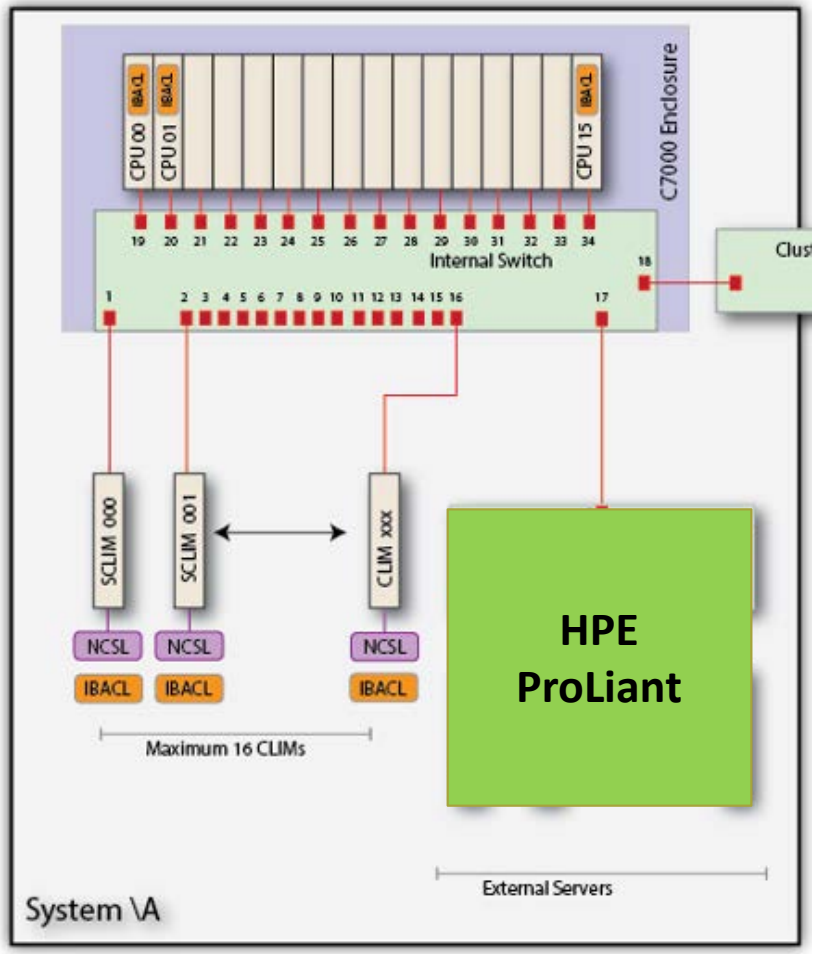
com.forte®

# Phase 2 of comForte Yuma POC - Context

> Now all on comForte Hardware
  - > **comForte owned and operated NS3 X1**
  - > **HPE ProLiant, RHEL Linux, Mellanox IB card**
  - > **Only a single InfiniBand cable, no Switch on "Linux end" of connection**

> Still with plenty of help from HPE folks
  - > **Direct contact with key developers**
  - > **Direct contact with HPE product management**

**Thank you much HPE!**



# Phase 2 of comForte Yuma POC - Context

> comForte Resources for Phase 2

- **comForte: Thomas Burg, various folks in sys admin NonStop and Linux**
- **Gemini: Richard Pope, Dave Cikra**

> Gemini Communications, Inc.

- www.geminic.com
- **No direct sales**
- **Several 'comm' products over the decades, some of them sold by comForte now**

# Phase 2 of comForte Yuma POC - Goals

> Compare InfiniBand with 1 Gbit TCP/IP
  > **Like all NS3 X1, comForte system does not have 10 Gbit Ethernet**
  > **Hence 10 Gbit could not be measured**
  > **Compare 1 Gbit Ethernet with InfiniBand**

> Re-measure some key data points for ‘moving of data’:
  > **Latency and throughput for ‘typical’ packet sizes**
  > **Maximum throughput using ‘optimal’ packet sizes**

> Can we do ‘FTP over InfiniBand’ and if so, how fast?

# Phase 2 of comForte Yuma POC - Disclaimer

> It has been a tight race to GTUG

- > **The comForte NS3 X1 system was delivered in October 2015**
- > **The Linux system was set up in January 2016**
- > **The missing InfiniBand cable was ordered in February 2016**
- > **InfiniBand was up and running in March 2016**

>*Please treat all number as preliminary. Things should only get better, but all numbers are the result of a POC, rather than benchmarks of a finished product*

# The comForte Yuma POC Phase 2

# Moving Data

better always on

com.forte®

# Moving data – model used

> For “POC Phase 1” (TBC Nov 2015) we used ‘*echo*’ approach
  > **Send some bytes of data**
  > **Send same packet size back**

> For “POC Phase 2” (GTUG April 2016) we used ‘*one way*’ approach
  > **Send some bytes of data**
  > **Send small packet (“acknowledgement”) back**

> Both models occur in real life, but we felt ‘one way’ is more common

# Moving data 16 KBytes – results

> ‘one way’ approach (see prior slide)
  > **16 KBytes = 16384 bytes data, 20 bytes “ack”**
  > **Data moves from NonStop to Linux**

| Transport over | Latency (microseconds) | MegaBytes/s |
|---|---|---|
| TCP/IP 1 Gbit Ethernet | 374 | 43 |
| InfiniBand | 11 | 1413 |
| **InfiniBand gain** | **x 34** | **x 32** |

# Moving data optimum packet size – results

> ‘one way’ approach

- **‘Optimal’ packet size chosen for InfiniBand and TCP/IP, “ack” still 20 bytes**
- **Data moves from NonStop to Linux**

| Transport over | Packet size | Chunk of data moved to measure real time [in GigaBytes] | Real time elapsed [in seconds] | Throughput [in MegaBytes/s] |
|---|---|---|---|---|
| TCP/IP 1 Gbit Ethernet | 262 144 | 10 | 97 | 102 |
| InfiniBand | 2 097 152 | 1024<br>[one TeraByte] | 176 | 5734 |
| **InfiniBand gain** | | | | **x 55** |

- **time to move one TeraByte over TCP/IP 1 Gbit Ethernet extrapolates to 9900 seconds**

# The comForte Yuma POC Phase 2

# Moving files from NonStop to Linux

better always on

com.forte®

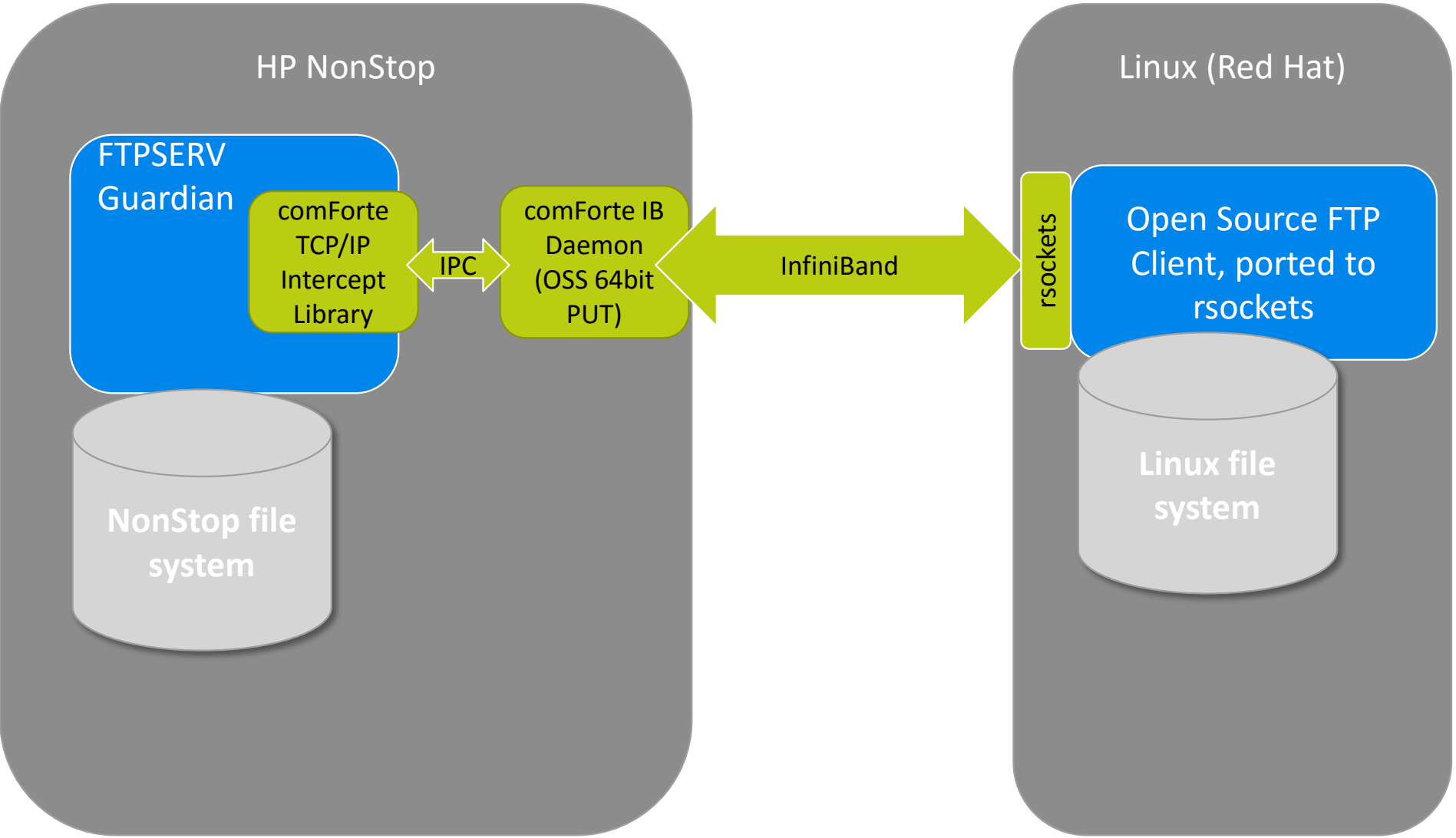
# ‘FTP’ over InfiniBand - introduction

- During POC phase 1 comForte and Gemini managed to connect NonStop FTPSERV with Linux open source FTP client
  - **No modifications to NonStop FTPSERV (!). Used comForte “TCP/IP to InfiniBand intercept framework” (see next slide)**
  - **Converted Linux open source FTP client to rsockets**

- FTP protocol is NOT ‘InfiniBand’ friendly

- During POC phase 2 we focused on speed measurements, hence we wrote test programs with direct file I/O on both ends

# comForte FTPSERV over IB POC (done for TBC 2015)




> This worked, but it needed some 'tricks'

> Performance was good, but not faster than 10 Gbit Ethernet, about 300 MB/s

> Works for Telnet as well ☺

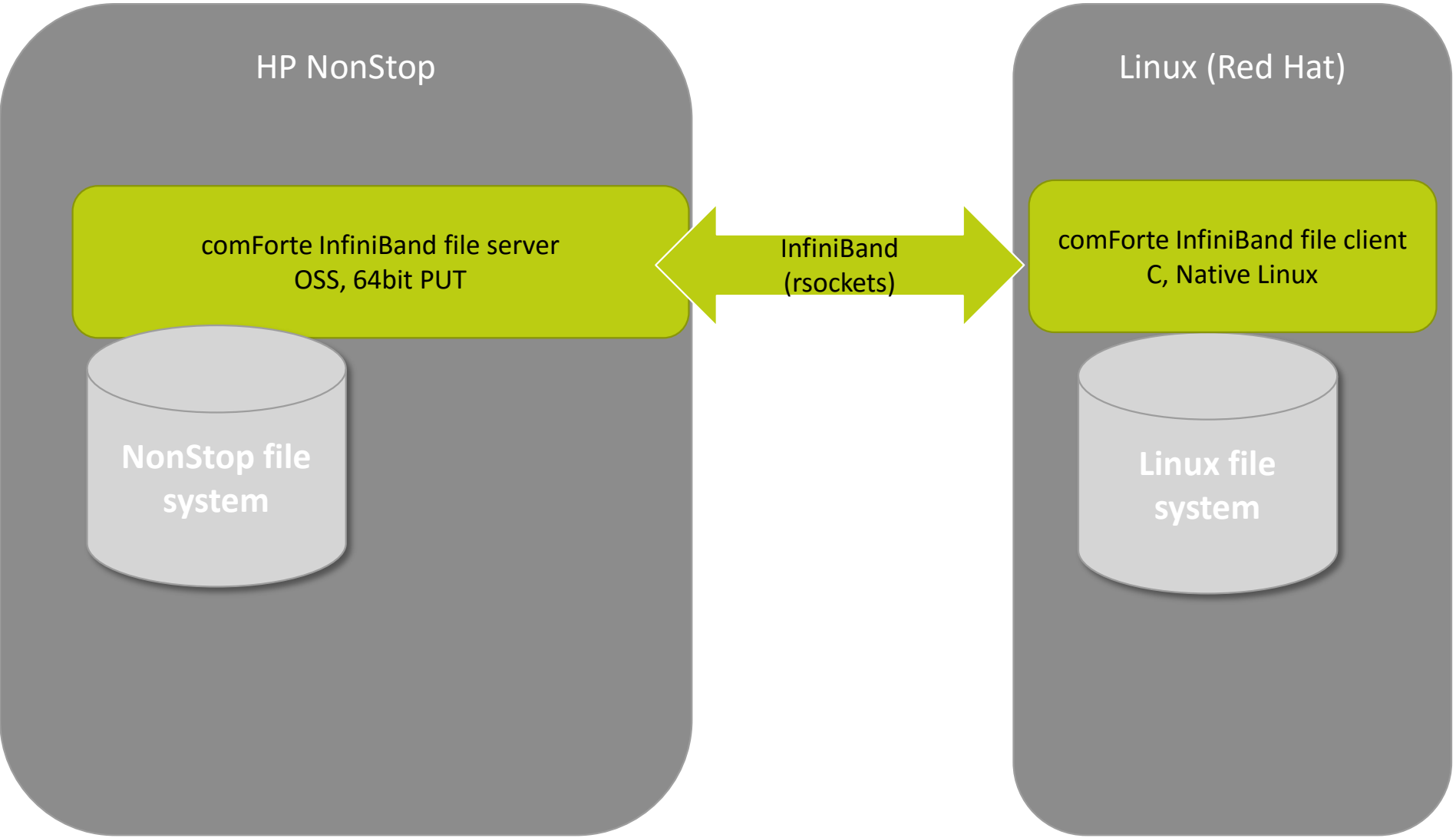
# ‘FTP’ over InfiniBand – changes for Phase 2 of POC

> No longer use FTP protocol at all

> Have comForte code on both ends

> Full control, no extra IPC between Guardian and OSS layer

# comForte ‘FTP’ over Infiniband April 2016



# FTP over TCP/IP, 1 Gbit Ethernet

> Single *file read* maxes out @ about 150 MByte/s [used test program for this]
> TCP/IP maxes out @ about 128 MByte/s
> FTP file transfers based on number of parallel transfers for a 1 GigaByte file from NonStop to Linux

| Number of parallel file transfers | Total time elapsed (seconds) | Throughput (Megabyte/s) |
|---|---|---|
| 1 | 19 | 53.9 |
| 2 | 28 | 73.1 |
| 4 | 45 | 91.0 |
| 8 | 75 | 109.2 |
| 16 | 147 | 111.5 |

# ‘FTP’ over Infiniband – POC results

> **InfiniBand has no real limit here ☺ [it is about 6 GByte/s]**
> **‘FTP’ file transfers based on number of parallel transfers, same file, but now over InfiniBand:**

| Number of parallel file transfers | Total time elapsed (seconds) | Throughput (Megabyte/s) |
|---|---|---|
| 1 | 10 | 102.4 |
| 2 | 10 | 204.8 |
| 4 | 10 | 409.6 |

> **Already moved from 111 MByte/s to 410 MByte/s → Nearly four times faster**
> **Limitations of file transfer speed are now:**
  > How effectively can we “scale out” File I/O read operation
  > This was measured on a two CPU NS3 X1

# Moving data from NonStop to Linux – testing the limits on Linux and InfiniBand

> Use ‘FTP over InfiniBand’ POC framework

> Do *not* do file read on NonStop, use test data created in memory

> Send data to Linux, flush to disk

> This measures

  > **Disk write speed on Linux**

  > **How well current comForte POC FTP over InfiniBand file server and client scale**

# Moving data from NonStop to Linux – testing the limits on Linux and InfiniBand

| Number of parallel sessions | Average speed of individual sessions (in MegaByte / s) | Throughput (Megabyte/s) |
|---|---|---|
| 1 | 622 | 622 |
| 4 | 550 | 2210 |
| 6 | 400 | 2460 |

> **Scales up nicely on a two CPU system with a single InfiniBand cable**

# What to make of ‘FTP over IB’ results

- **comForte can move data real fast from NonStop to Linux**
  - 6 GigaBytes per second seems doable on a fully scaled out NS7 X1
  - This includes flushing the data to Linux Disk

- **Potential use cases (???):**
  - Fast replacement for FTP
  - Data replication
  - Big data
  - Backup

# The comForte Yuma POC Phase 2

# Other observations

better always on

com.forte®

# Other observations during POC

- **Setting up InfiniBand hardware on NonStop and Linux is new to sysadmin folks (both on Hardware and Software level)**

- **InfiniBand rsockets interface is straightforward to code, both on NonStop and Linux**
- **InfiniBand Low level verb interface is NOT straightforward to code**
  - Did not get beyond very early POC code but making progress

- **InfiniBand and rsockets are rock solid both on NonStop and RHEL Linux**

- **rsockets is only available from OSS PUT64 (not available under Guardian!). That's why comForte built a plug-compatible sockets DLL for Guardian socket apps (like CSL, FTPSERV, anything using TCP/IP under Guardian)**

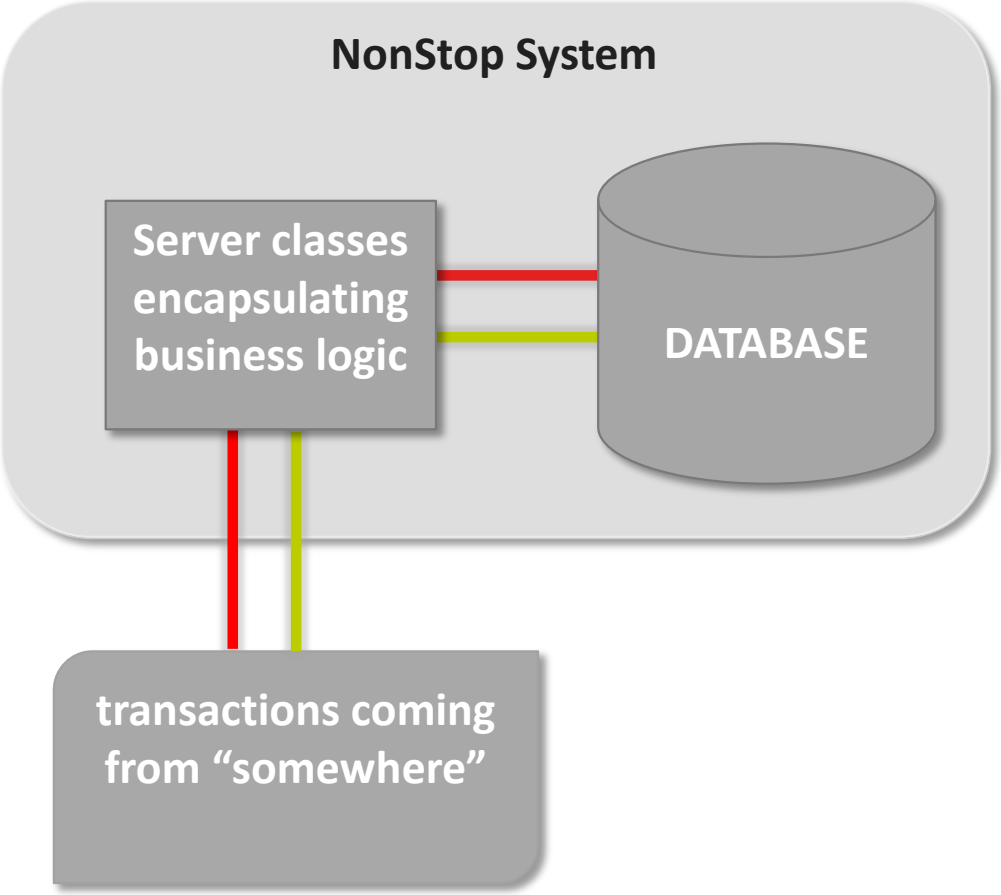
- **HPE NonStop InfiniBand team very competent and helpful**

# A business case and technical vision for „Hybrid NonStop“

better always on

com.forte®

# Cloud Business Case “Looking versus Booking”: Many NonStop systems as of today



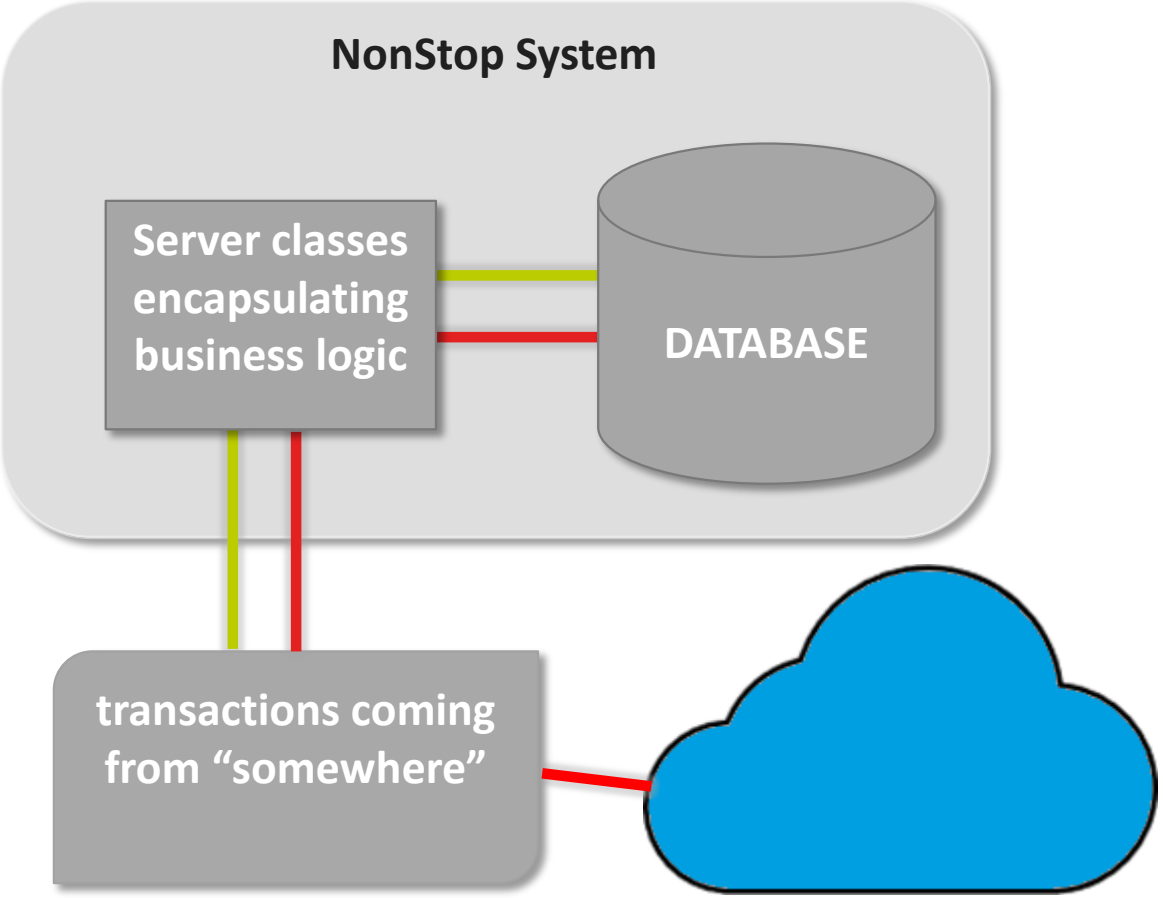

*Looking* and *Booking* traffic (typical use case for multiple NonStop customers in travel section):

**Looking is stateless, 95+% of traffic**
By nature of transaction, can be hosted in cloud or on commodity platform

**Booking is transactional**
By nature of data, you don’t want to lose it and it also has “state” (*ACID*) – run it on NonStop

Similar two-types-of-transaction logic applies to stock exchanges, potentially other verticals (Base24 !?)

# Cloud Business Case “Looking versus Booking”: The high level requirement/vision

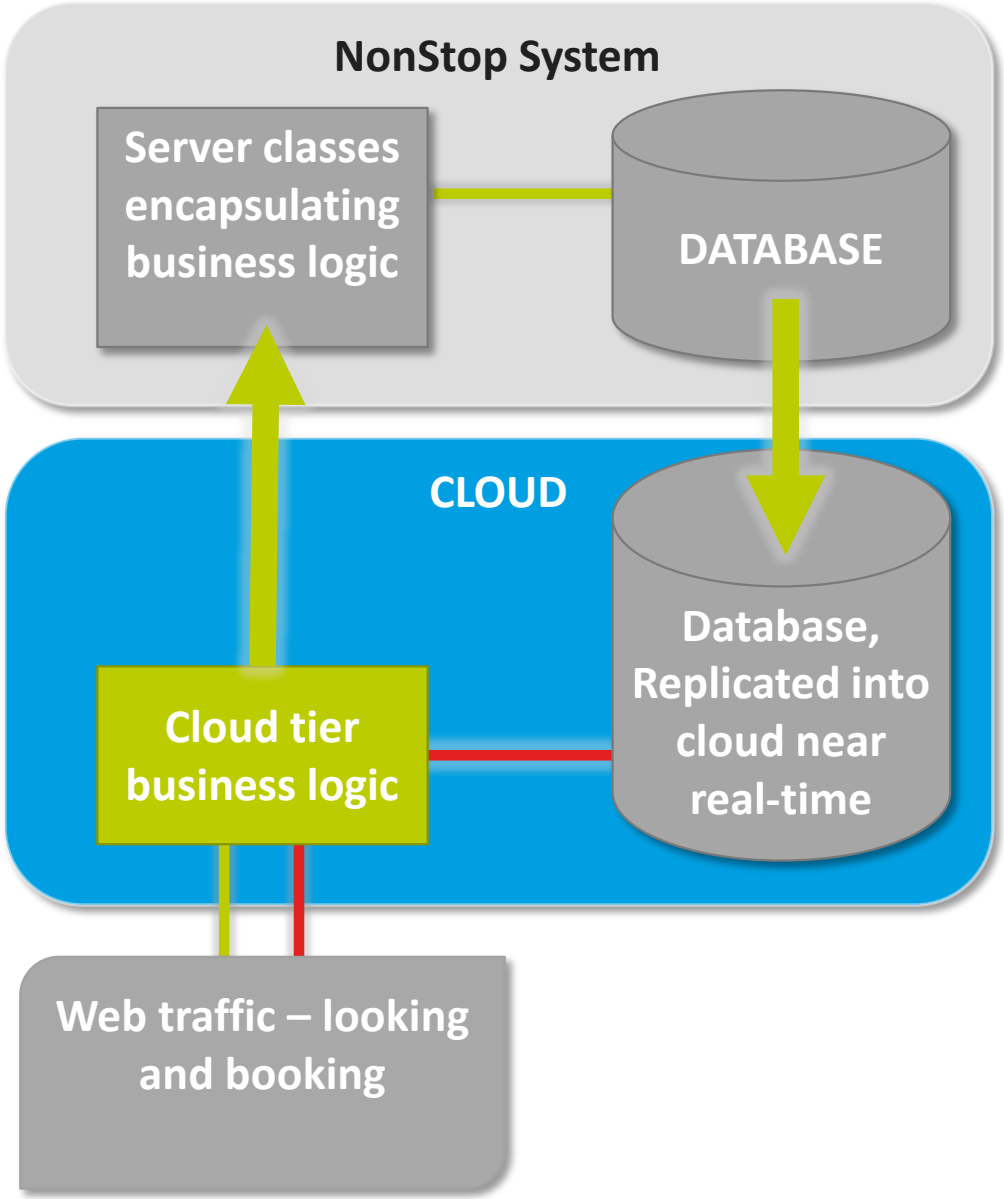



**Looking does not hit NonStop at all**

**… and is handled in the cloud (public or private)**

**… but how to move ‘state’ (database) to cloud???**

# Cloud Business Case “Looking versus Booking” – InfiniBand and NonStop Hybrid vision




**Looking indeed handled in cloud**
transactions do not hit NonStop (business tier knows it is *Looking* and hence simply uses local DB copy)

**Cloud tier sends Booking transactions to NonStop, via Infiniband** (again, business logic sees this is *Booking*, hence switches to NonStop)

**Fast replication via InfiniBand** enables (one-way, “read only”, near real-time) replication to multiple Linux boxes in parallel with low latency and low CPU overhead

# The comForte Yuma Product Vision

- CSL/Infiniband
- Become *the* company for IB-enabling applications and middleware products
- Work with ISVs, end users

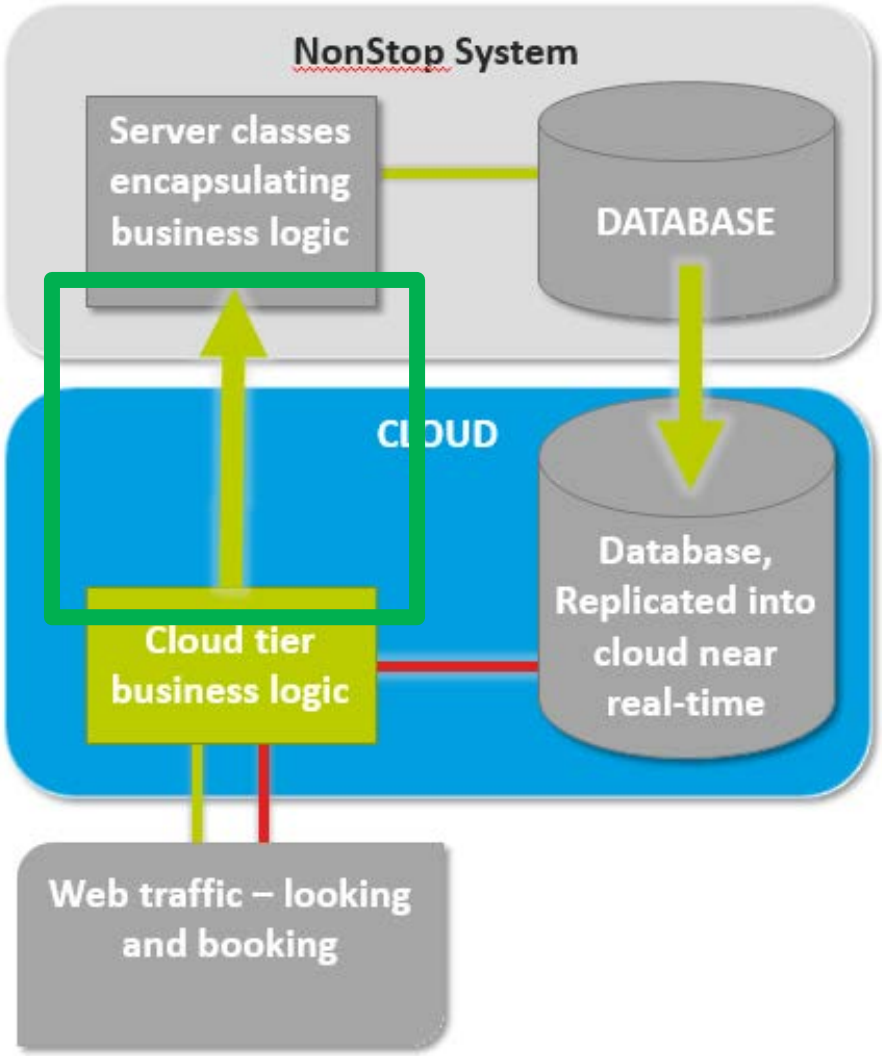
better always on

com.forte®

# CSL/InfiniBand

> Covers “left half” of InfiniBand Hybrid vision

> Available very soon…



# CSL/InfiniBand

> A very natural extension of the CSL product

> A new option CSL/InfiniBand
  > **First release will provide C/C++ API on Linux**
  > **To be announced @ GTUG Berlin, again at TBC 2016**
  > **EAP-ready October 2016**

> Come to comForte presentation or talk to us to find out more

| Tuesday, 19 April 2016 | Aquamarin |
|---|---|
| 17.15 - 18.15 | **Hybrid is today!**<br>**InfiniBand-Enable your applications**<br>*Thomas Burg, comForte21* E |

# The broader comForte Yuma framework

> Can InfiniBand-enable *any* existing application on HPE NonStop
  - **Without application changes (!)**
  - **Just like SecurData and CSL – it is a *framework***

> Existing application/middleware on NonStop
  - **InfiniBand will boost performance**
  - **With comForte experience from POC and framework to be announced @ TBC 2016, Middleware/application vendors can focus on their features, comForte takes care of InfiniBand details**

> Customer/partner needs to do work on Linux himself
  - **Rather easy via rsockets approach**
  - **comForte can provide proxy (speed to be confirmed)**
  - **comForte can help**

**> comForte vision: Become THE player in "InfiniBand low level coding"**

# The broader comForte Yuma framework (contd.)

- Whom does this help moving from TCP/IP to InfiniBand?
  - **NonStop ISVs**
  - **Software houses with their own applications**
  - **NonStop users with their own applications**

- Interested?
  - **Come talk to comForte**

# Summary, Q&A

better always on

com.forte®

# Summary, Q&A

- HPE NonStop is now InfiniBand enabled to connect to HPE ProLiant Server running RHEL
- InfiniBand is extremely fast
- Now that HPE has created an environment that we can build on with InfiniBand, comForte has several products which can be used in the “Hybrid space”
  - **CSL/InfiniBand**
  - **InfiniBand enabling framework**
  - **maRunga/InfiniBand (?)**
- **Time to start moving to Hybrid!?**

“full roundtrip time”
= 11 micro seconds

= 90.000 TPS

= 34 x faster

3 min to move a TeraByte

= 55 times faster

**THANK YOU !**

**Questions?**