Hewlett Packard
Enterprise

NonStop Technical Boot Camp 2023
TBC23-TB59 How the NonStop Solution Excels in the Industry

Roland Lemoine, Product Manager

September 2023

HPE Partner and Customer Use Only

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.

Agenda

Horizontal scale, shared nothing and no replicas: This is the way!

What makes NonStop availability superior

NonStop, the best cluster you can find

The NonStop unique architecture

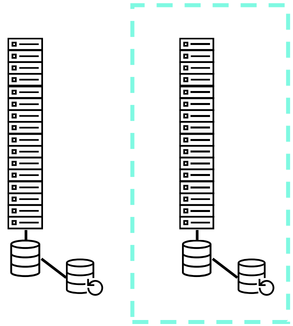
AL2

Big fault zones

High failover time

HA fully relies on DR site switch over

Scale up



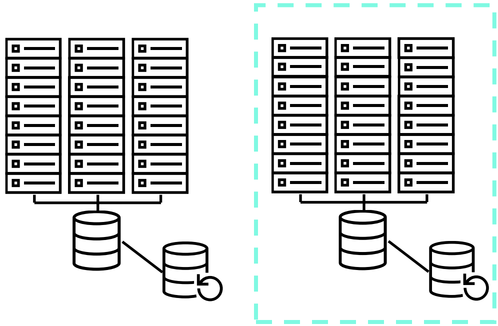
AL3 (some visible failures)

Limited scale out

Reduced failover time (30s)

Shared disk requires cache-coherency in each node

Scale out / Shared disk



AL4 (no visible failures) (*)

No performance bottleneck

Big data/NoSQL

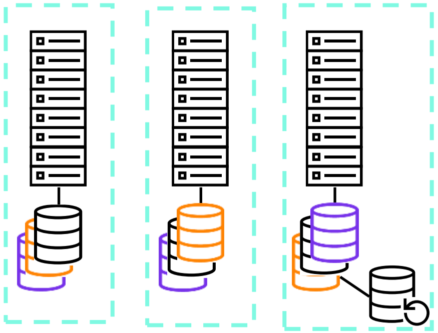
Eventual consistency

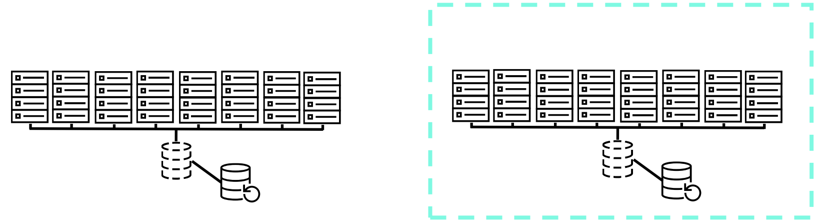
High parallelism for OLTP and analytics

DR site efficiencies

State of the art and cost effective

Scale out / Shared nothing

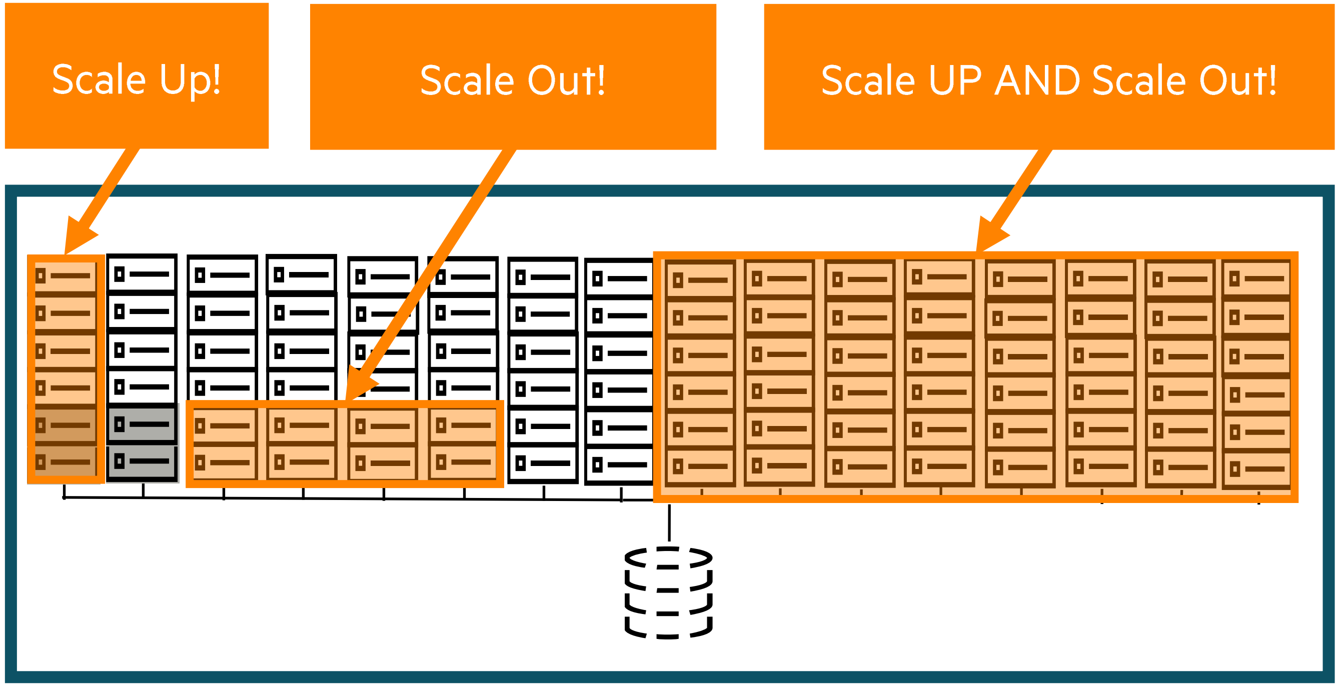




NonStop is scale out, shared nothing, without replicas

(*) Source: IDC, Jul 2020, Doc #US46640020, Worldwide AL4 Server Market Shares, 2019.

Multi-dimensional and linear scale with a single architecture



A NonStop system

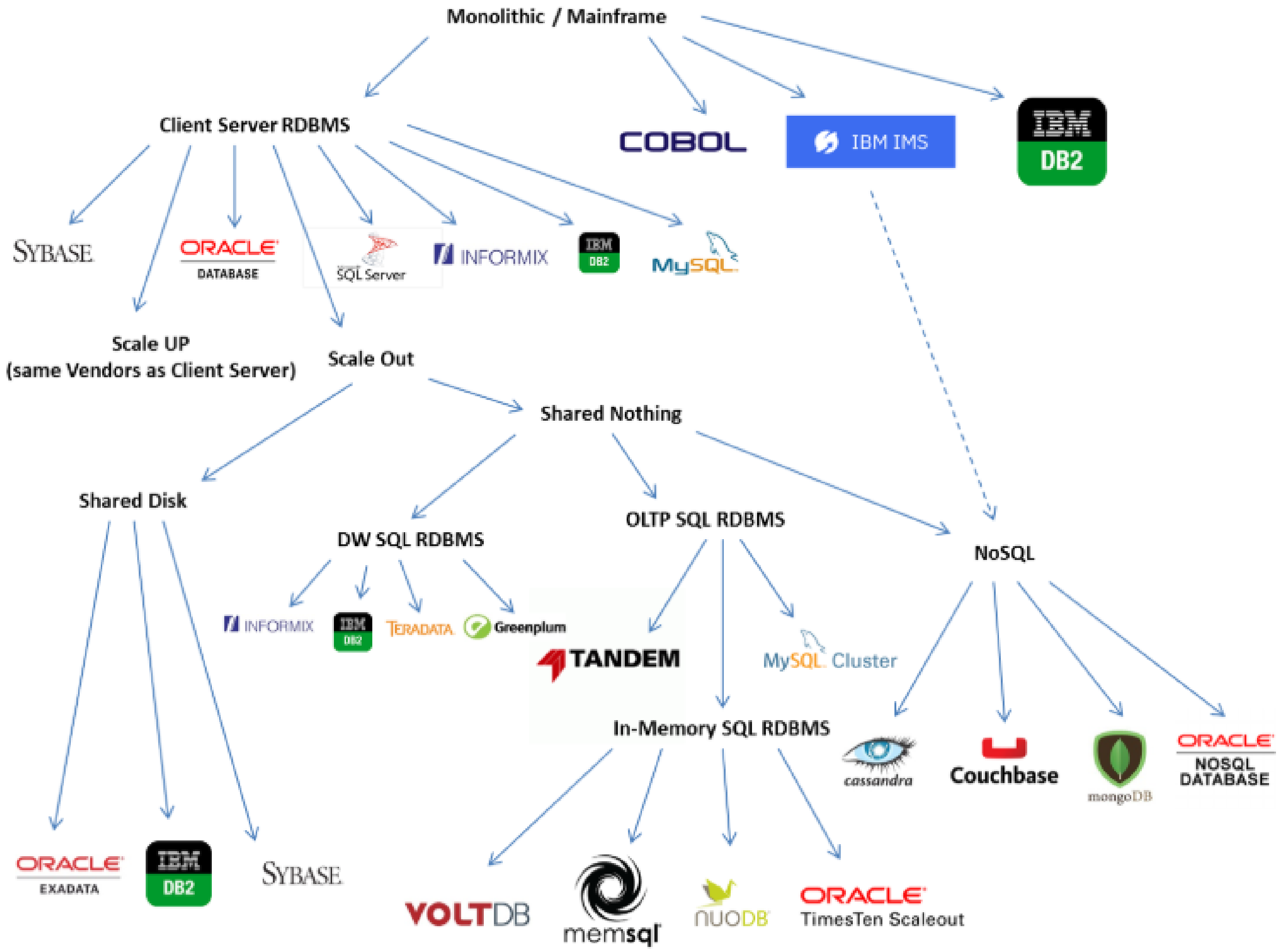
- Nodes can be added without stopping the application
- Cores can be turned on and off without stopping the application
- Scale out to 16 nodes (CPUs in NonStop terms)
- Each node scale up to 6 cores
- Disks are virtualized and visible from all nodes

NonStop uses a shared nothing architecture to scale beyond a few nodes efficiently. No architecture change when moving from development to production. Added capacity fully translates in added throughput

Databases that adopted scale out and shared nothing

Going back to the 70's IBM, Oracle and Ingres, Sybase and Informix where the early vendors introducing the concept of a relational database accessed using the SQL language

- Scale up vs scale out was the first split of 2 very different architectures
- Under scale out another split occurred based on using shared disks between the nodes or not physically sharing them
- NonStop SQL is a scale out, shared nothing architecture



Monolithic / Mainframe

Client Server RDBMS

Scale UP
(same Vendors as Client Server)

Scale Out

Shared Nothing

Shared Disk

DW SQL RDBMS

OLTP SQL RDBMS

NoSQL

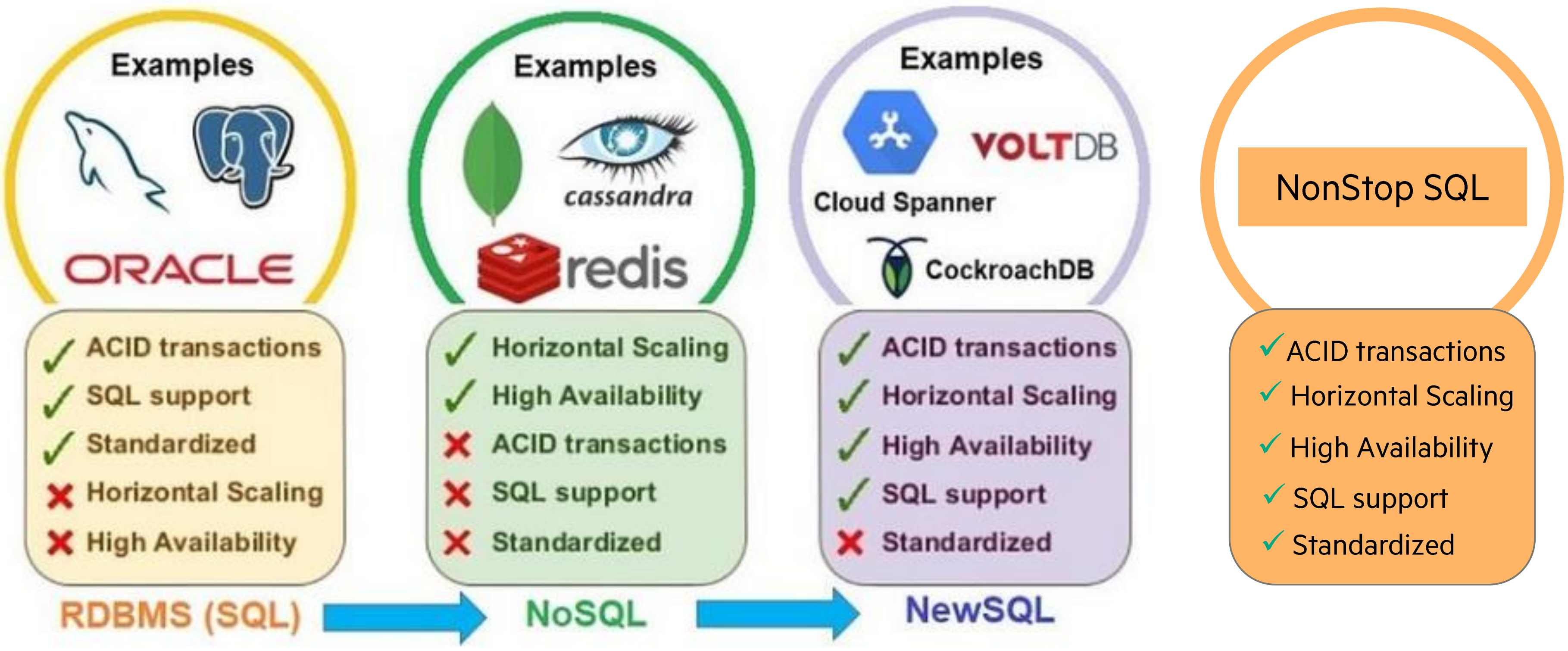
In-Memory SQL RDBMS

* https://blogs.oracle.com/timesten/the-evolution-of-db-architectures Oct 2017

HPE Partner and Customer Use Only

Scale out but with various levels of success

- NoSQL adopted scale out to address Bigdata higher scale requirements
- NewSQL to mitigate lack of transactions and SQL of NoSQL
- NonStop SQL has no trade-off for transactions yet scales horizontally



Examples

ORACLE

✓ ACID transactions
✓ SQL support
✓ Standardized
✗ Horizontal Scaling
✗ High Availability

RDBMS (SQL)

Examples

cassandra

redis

✓ Horizontal Scaling
✓ High Availability
✗ ACID transactions
✗ SQL support
✗ Standardized

NoSQL

Examples

Cloud Spanner

VOLTDB

CockroachDB

✓ ACID transactions
✓ Horizontal Scaling
✓ High Availability
✓ SQL support
✗ Standardized

NewSQL

NonStop SQL

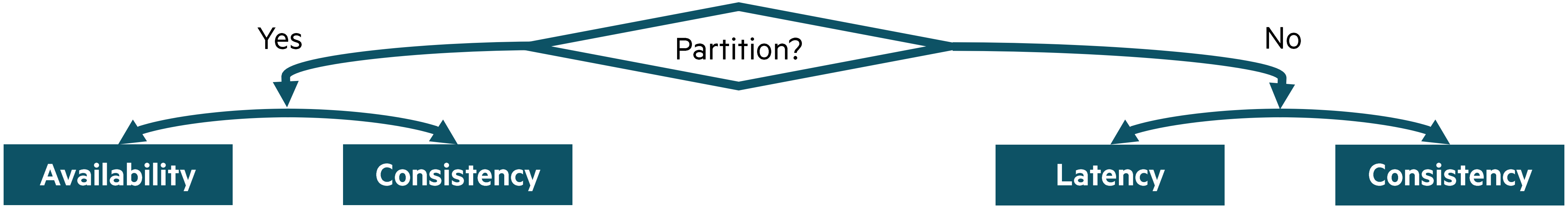
✓ ACID transactions
✓ Horizontal Scaling
✓ High Availability
✓ SQL support
✓ Standardized

* https://medium.com/rabiprasadpadhy/google-spanner-a-newsql-journey-or-beginning-of-the-end-of-the-nosql-era-3785be8e5c38

More trade-offs when using replicas to achieve high availability

PACELC theorem – a revised and more complete version of the CAP theorem

in case of network partitioning (P) in a distributed computer system, one has to choose between availability (A) and consistency (C) (as per the CAP theorem), but else (E), even when the system is running normally in the absence of partitions, one has to choose between latency (L) and consistency (C)



Trade-off in network failure scenarios
If you favor consistency, wait for end of failure
If you favor availability, you may get stale data

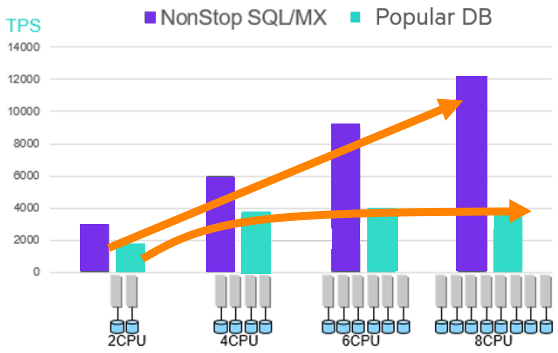
Trade-off in normal processing scenario
If you favor consistency, wait for end of all writes
If you favor latency, you may get stale data

- For example Kafka uses “min.insync.replicas” so it is up to the end-user to decide how replicas are synchronized based on factors such as “multi-zone or multi region deployment, latency within or outside a region, consistency within or outside a region (https://www.ibm.com/cloud/architecture/architecture/practices/strategies-for-kafka-reliability/)
- Kubernetes “etcd” (Kubernetes metadata stored in the control plane) database
- MongoDB in failure scenario favors consistency, but writes will be suspended until a new leader is elected (~12 seconds)

Scalability and resilience using multi-dimensional linear scale

- **Linear scalability**
  - Unlimited expandability and capacity using “shared nothing”
  - Combine vertical (cores) and horizontal scale (nodes) for optimal performance
  - Grow from prototype to mission critical without re-architecting or rewriting the application
  - No architectural bottleneck when adding capacity
  - No increase in latency for high availability purposes
  - Add cores or nodes without stopping the application

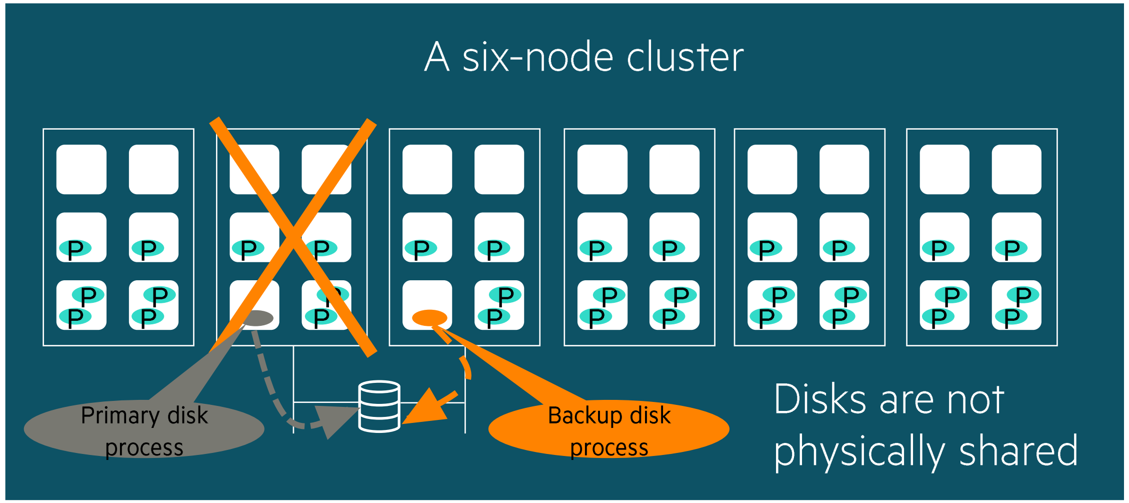
Adding nodes translates to 98% increase in throughput



98.3%

Linear Scalability

Better design drives better outcomes


A six-node cluster

Disks are not physically shared

Only one node owns the disk at a time (shared nothing), other nodes ship the I/O to this node via message passing. If the node fails, the backup process becomes the owner without visible impact to the application. Combining “shared nothing” and “process pair” for storage access relieves the NonStop architecture to use replicas or complex distributed locking.

Using replicas as it is generally done in the rest of the industry raises many challenges and limitations such as split brain, CAP (*) theorem restrictions, need for consensus algorithm, double writes or ghost I/O and overall solution limited scalability

* CAP stands for consistency, availability, and partition tolerance

HPE Partner and Customer Use Only

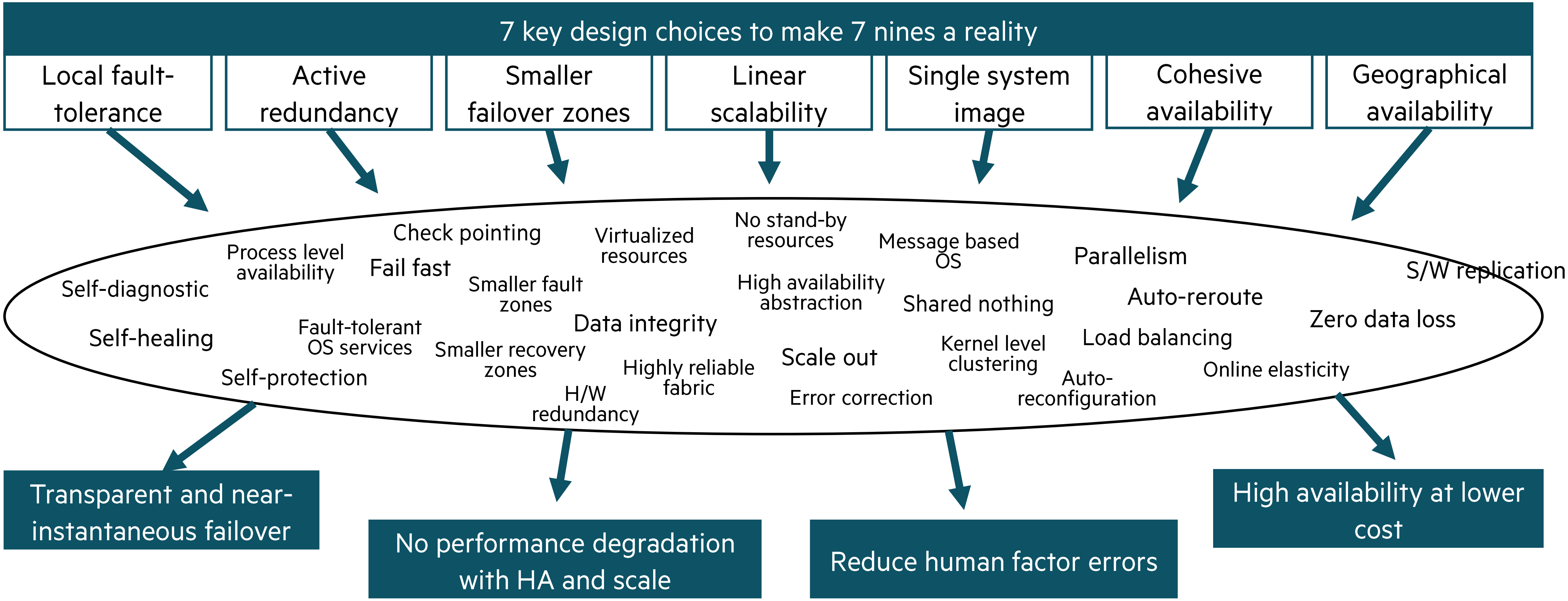
Agenda

Horizontal scale, shared nothing and no replicas: This is the way!

What makes NonStop availability superior

NonStop the best cluster you can find

What makes NonStop availability superior



7 key design choices to make 7 nines a reality

- Local fault-tolerance
- Active redundancy
- Smaller failover zones
- Linear scalability
- Single system image
- Cohesive availability
- Geographical availability

Self-diagnostic

Self-healing

Process level availability

Fault-tolerant OS services

Self-protection

Check pointing

Fail fast

Smaller fault zones

Smaller recovery zones

Virtualized resources

Data integrity

H/W redundancy

Highly reliable fabric

No stand-by resources

High availability abstraction

Scale out

Error correction

Message based OS

Shared nothing

Kernel level clustering

Parallelism

Auto-reroute

Load balancing

Auto-reconfiguration

Online elasticity

S/W replication

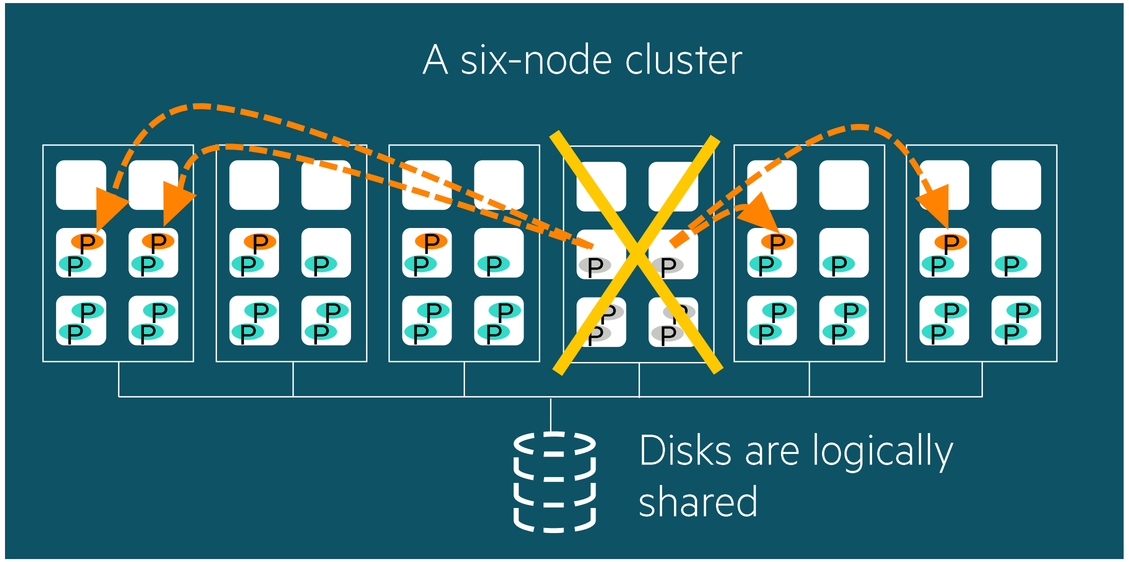
Zero data loss

- Transparent and near-instantaneous failover
- No performance degradation with HA and scale
- Reduce human factor errors
- High availability at lower cost

Ultimate data availability and integrity

- **No data loss or corruption**
  - Millisecond takeover instead of failover
  - Strong consistency for read/writes at scale
  - Fault tolerant engine for ACID transactions
  - Transaction aware replication and validation at geographical scale including zero data loss for disaster recovery
- **No down time**
  - Fault-tolerance via a shared nothing, full system cluster
  - Fault tolerant OS services and database
  - Active redundancy using process check points
  - No visible failures for the application
  - No single point of failure reference architecture
  - Immediate, automated reroute of workloads in case of failures
  - Zero downtime data migration and system updates
  - Rated AL4 by IDC (*)

A self-healing proven architecture



In case of a node failure, tasks are migrated in milliseconds vs multi-seconds OS/DB failover on other platforms

Redistributing one node's workload (16% of overall) over the remaining 5 nodes means they only need to take on an additional 3% charge, preserving response times and scale

* An IDC document (#US46640020), July 2020

Agenda

Horizontal scale, shared nothing and no replicas: This is the way!

What makes NonStop availability superior

NonStop the best cluster you can find

“Design isn’t crafting a beautiful textured button with breathtaking animation. It’s figuring out if there’s a way to get rid of the button altogether”

- Edward Tufte

“Single system image is a computing paradigm where a number of distributed computing resources are aggregated and presented via an interface that maintains the illusion of interaction with a single system”

“Kernel-level SSI seeks to diminish the effort required to utilize distributed computing resources through transparent aggregation”

- Philip Healy, Theo Lynn, Enda Barrett, John P. Morrison
https://www.researchgate.net/publication/295253720_Single_system_image_A_survey

The NonStop cluster characteristics

Single System Image

- No regression compared to the SMP model for applications to benefit of parallel processing
  (auto processor assignment)
- No need to implement an external load balancer
- No need to develop cluster aware versions of the manageability tools
- No need for cluster management s/w
- System administration simplified
- System security implementation simplified
- No need for applications to implement their own clustering

Full system cluster

- Kernel SSI means the cluster boundaries are the same for any software installed on the system
- All users: admin, dev, end-users have the same view of a single system
- High availability turned on by default for the whole s/w stack
- No need to install s/w on each node
- No need to assign IP addresses and ports for clustering intra-system exchanges
- Middleware and networking layers automatically take advantage of the cluster and SSI

High value

- scale, availability, load balancing and SSI built-in (not configured)

No trade-offs

- No PACELC or CAP theorem trade-offs
- No consensus algorithm constraints
- No need to setup replicas

Industry clustering solutions: Neither Kernel SSI neither full system cluster

The SSI regression

A cluster that does not implement a full system Single System Image (SSI) is a system that introduces a regression compared to SMP systems

No OS standard

OS clustering features are different even between Linux distribution. This means middleware have to re-invent their own and OS clustering cannot become mainstream

No cluster blueprint

As seen previously with replicas trade-offs, clustering requires a lot of configuration decisions

Partial clustering & SSI

Beowulf cluster (ssh and NFS)

Veritas, Lustre, LVS, Linux Pacemaker

HPE Serviceguard, IBM PowerHA

VMware

Sample middleware clustering

Oracle RAC, SQL Server

Kafka

Zookeeper

Kubernetes EKS, GKS,..

OpenShift

…

Manually assembled clusters

OpenStack

Consensus algorithms

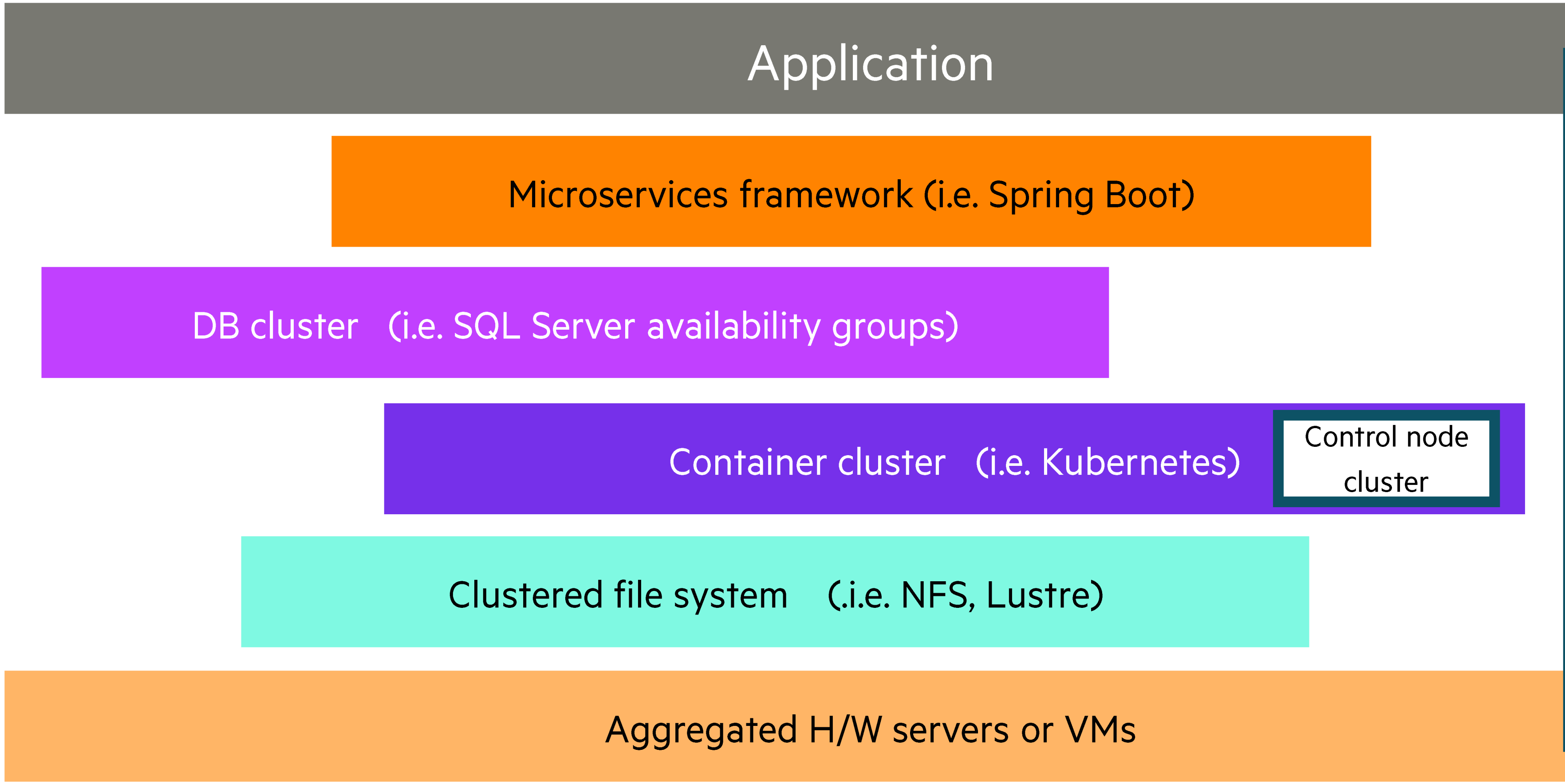
Read replicas

PostgreSQL/NFS

Redundant efforts

Too complex & high risk

The outcome of non-SSI and partial cluster



Application

Microservices framework (i.e. Spring Boot)

DB cluster (i.e. SQL Server availability groups)

Container cluster (i.e. Kubernetes)

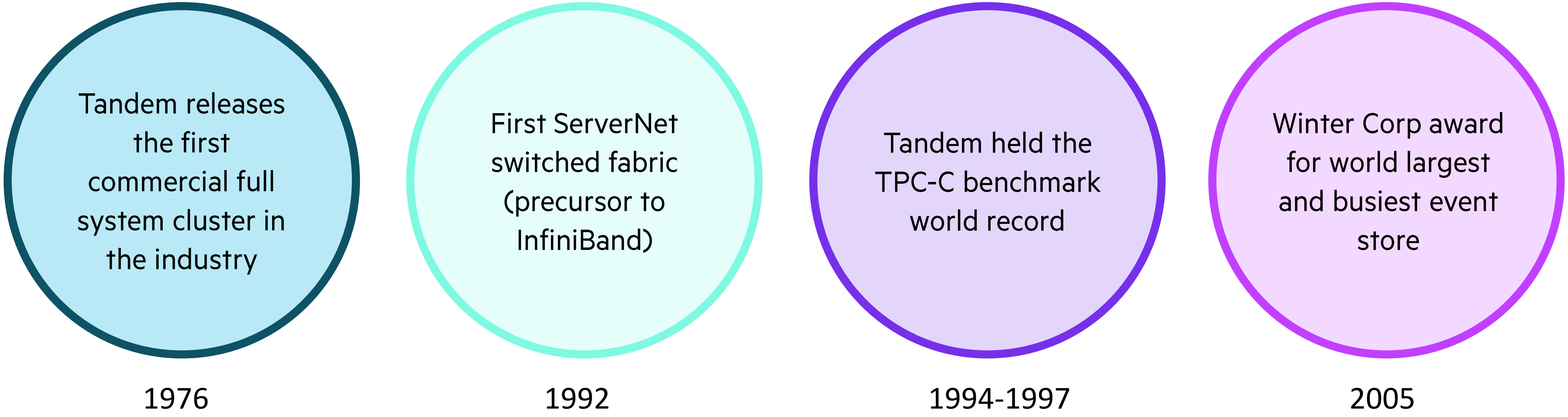
Control node cluster

Clustered file system (.i.e. NFS, Lustre)

Aggregated H/W servers or VMs

- Each layer uses its own clustering:
  - Different consensus
  - Different failover times
  - Different skills
- May have different boundaries
- Different users have different views
- Often need clustering for their control plane in addition to the cluster itself

NonStop cluster industry recognition



Tandem releases the first commercial full system cluster in the industry

1976

First ServerNet switched fabric (precursor to InfiniBand)

1992

Tandem held the TPC-C benchmark world record

1994-1997

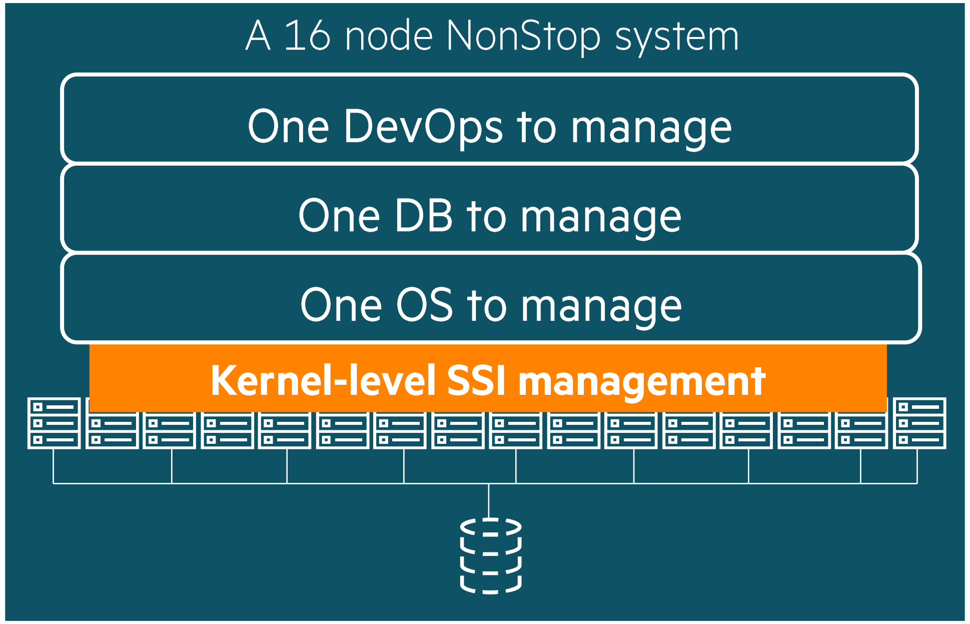
Winter Corp award for world largest and busiest event store

2005

Proof points of a leading and successful architecture

Reduced risk, effort and cost - part 1

- **Full-system cluster & Single System Image**
  - Simpler for the application, relieved of clustering efforts
  - Simpler for administration to manage a single system
  - Simpler to secure a single system
  - Only a small team required to manage NonStop



A 16 node NonStop system

One DevOps to manage

One DB to manage

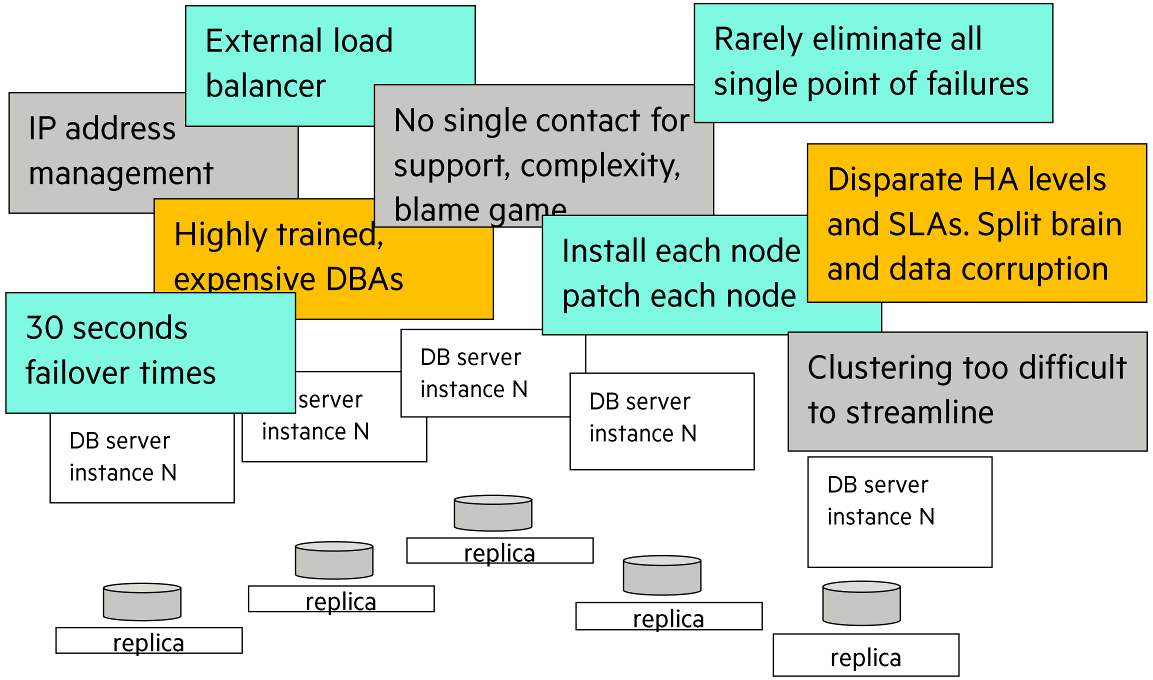
One OS to manage

Kernel-level SSI management

Kernel-level SSI management is the most desirable solution to manage clusters (*)

Avoid unnecessary complexity

Other “assembled” clusters



External load balancer

Rarely eliminate all single point of failures

IP address management

No single contact for support, complexity, blame game

Disparate HA levels and SLAs. Split brain and data corruption

Highly trained, expensive DBAs

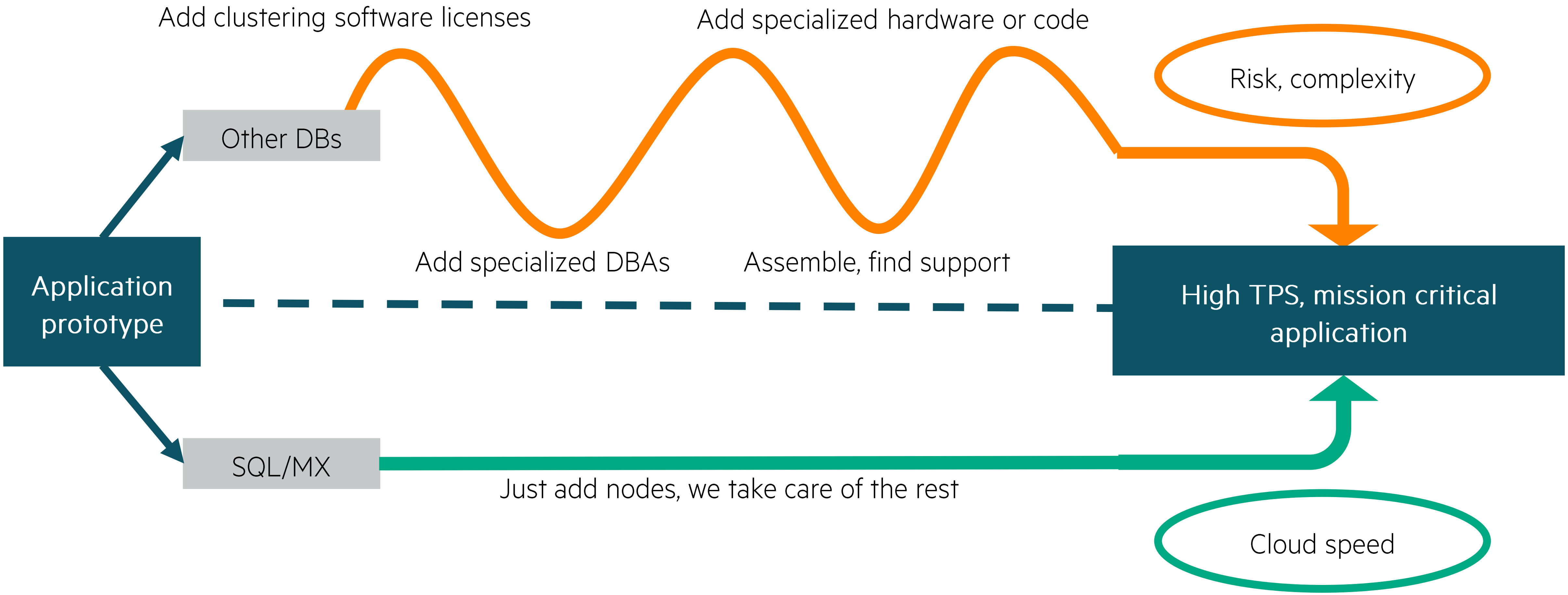
Install each node patch each node

30 seconds failover times

Clustering too difficult to streamline

(*) Gregory F. Pfister “In search of clusters” (1998)

Less cost, less risk for building mission critical applications

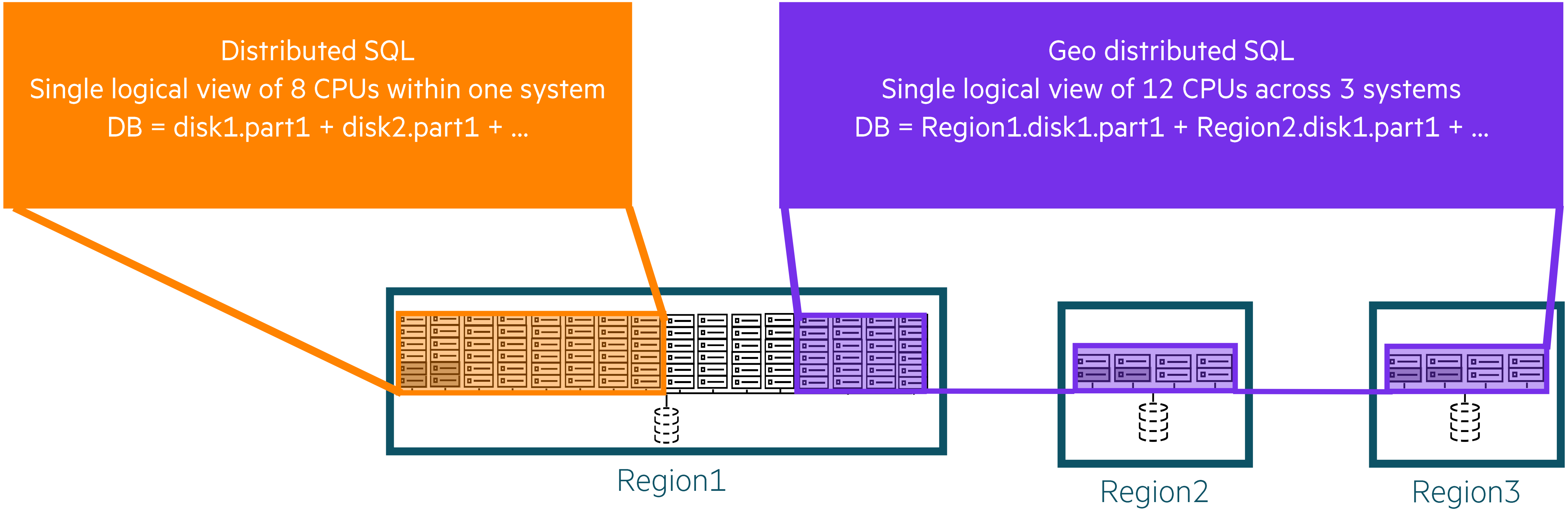

Add clustering software licenses
Add specialized hardware or code
Risk, complexity
Other DBs
Application prototype
Add specialized DBAs
Assemble, find support
High TPS, mission critical application
SQL/MX
Just add nodes, we take care of the rest
Cloud speed

[Geo-]distributed Database

What defines a SQL distributed database is that the entire database cluster looks like a single logical database to the application. For NonStop this is true within one system (one cluster) or multiple systems (supercluster).

No replication involved, NS SQL leverages the fault-tolerant clustered file system and global transaction engine

Effectively SQL/MX is a Globally distributed SQL using horizontal scaling and multi-shard ACID



Distributed SQL
Single logical view of 8 CPUs within one system
DB = disk1.part1 + disk2.part1 + …

Geo distributed SQL
Single logical view of 12 CPUs across 3 systems
DB = Region1.disk1.part1 + Region2.disk1.part1 + …

Region1

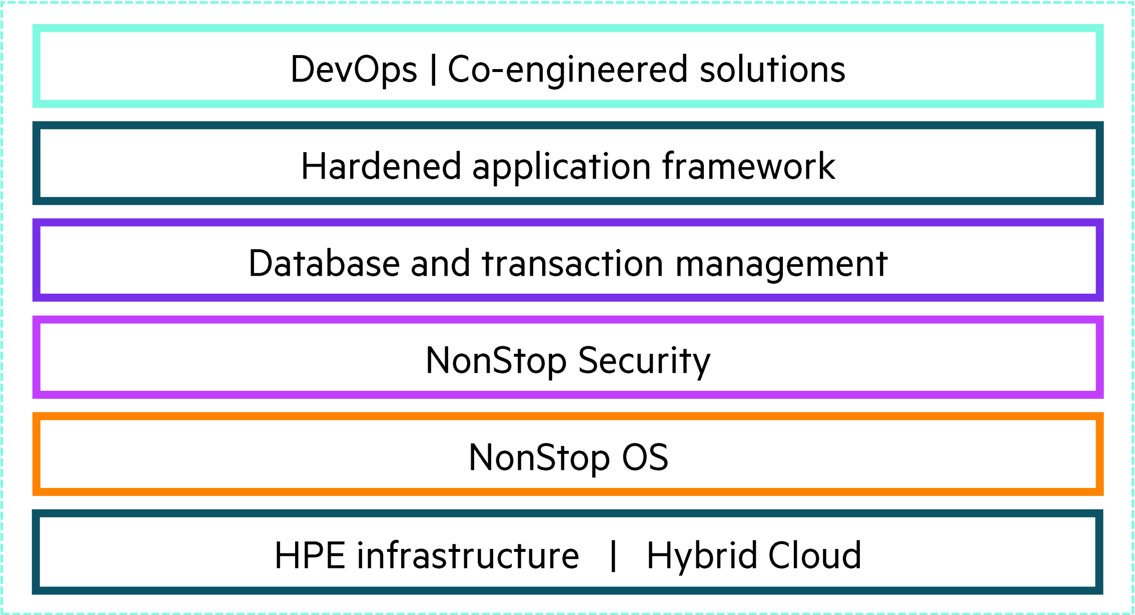
Region2

Region3

Reduced risk, effort and cost - part 2

- **Full & Integrated s/w stack**
  - Single vendor certification, sale, commitment and support
  - End-to-end security
  - Cohesive SLAs, security and compliance
  - Co-engineered architecture and performance optimizations
  - Make the most of the platform potential
- **HPE global excellence, reach, support and services at scale for the full stack**
  - Certified and highest security infrastructure
  - Global skills set with a Mission Critical culture
    - Solutions Architects
    - Advanced Technology Center
    - Support 24x7 and follow-the-sun
    - Professional Services
  - HPE Managed Services
  - HPE GreenLake consumption and co-location
  - Long term protected investment
  - 50 years of experience

The peace of mind



DevOps | Co-engineered solutions

Hardened application framework

Database and transaction management

NonStop Security

NonStop OS

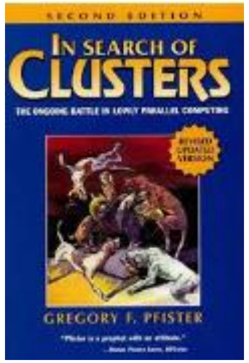
HPE infrastructure | Hybrid Cloud

No integration effort for the customer

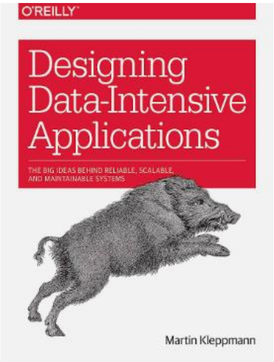
Reduced risk of a component lacking support or security

Recommended reading

- In search of clusters (Gregory F. Pfister)
  - Discuss cluster vs SMP
  - Full system clusters such as Tandem, OpenVMS and Parallel Sysplex
  - Discuss shared nothing, SSI



- Designing Data intensive Applications (Martin Kleppmann)
  - Discuss trade-offs associated with replicas architectures
    - Leader and followers, multi-leader, leader-less replicas, etc.
    - Consistency and consensus


O'REILLY
Designing Data-Intensive Applications
THE BIG IDEAS BEHIND RELIABLE, SCALABLE, AND MAINTAINABLE SYSTEMS
Martin Kleppmann

Conclusion

The NonStop architecture, scale out, shared nothing, no replicas is the best for better scale and availability

NonStop includes a much wider and advanced set of availability features

The NonStop cluster is a full system cluster with kernel level Single System Image that simplifies application development, administration and security while reducing risks

NonStop Partnership– It's a Beautiful Thing!

4techsoftware · abat · absyss · ACI UNIVERSAL PAYMENTS · ADVANCED SYSTEMS CONCEPTS, INC. · amdocs · Ascert · asseco · axway · BHMI · BlackWood Systems, Inc. · bmc

Bowden Systems Inc. · BPC · BrightStrand INTERNATIONAL MISSION CRITICAL SERVICES · ca technologies A Broadcom Company · CAIL · Canam Software XML and reporting solutions! · Capgemini · CAPSYS · CARR&SCOTT SOFTWARE INCORPORATED · Thycotic Centrify · CGI · CIRCA · DATA443

comforte · concerto Payment Systems Architects · connect Your Independent Hewlett Packard Enterprise Technology User Community · continuous NonStop IT Solutions · Cornerstone SOFTWARE · crossflo systems · CRYSTAL POINT · CS software · CSP · Daiwa Institute of Research

DXC.technology · ePROTEA FINEXUS · esq · ETINET · EUROMETRIC · eztips · FIS · FSS A Payment Systems Company · GEMINI Innovation & Leadership · GRAVIC HPE Shadowbase · greenHouse

IBM · IDELJI · IDMWORKS · IKAN · INETCO Every transaction tells a story · infrasoft · intel · INTELLIGENT WAVE INC. · ir · ITP SW Systeme · Japan Research Institute · Jit2000

KAUFFMANN ENTERPRISE · KCS (주)케이씨에스 · Knightcraft Technology Information Security Specialists · KSS 京信システムサービス Kyoshin System Service Co., Ltd. · lusis Payments · m2systems Transacting data, transforming companies. · MICRO FOCUS · MVI TECHNOLOGIES · NCI Network Concepts, Inc. · netweave INTEGRATED SOLUTIONS

Nexbridge · NonStopInsider · nti · NuWave MISSION-CRITICAL ENTERPRISE INTEGRATION · ODYSSEY INFORMATION SERVICES · OmniPayments · ORACLE · OSS Nokalva · PAXUS · PAYX · Pragnya solutions

PrimeCode Services Inc. · Progress formerly ipswitch · PYALLA Technologies · Qlik · QSA · River Rock Software · Rocket · Scipar, inc. · SMX I.T. SOLUTIONS PROVIDER Services & Consulting, Inc. · SI SOFTWARE INTEGRATORS

striim · Symmetryx · 凌群電腦 THE SYSCOM GROUP · TANDsoft NonStop Interception & Availability Solutions · tcm · THALES · TIC TICSOFTWARE NonStop Innovation · TIS TIS INTEC Group · BANKBOTTLENECKS Transaction Design, Inc. · TriCom · TrueByte

TRUSTED SECURITY SOLUTIONS · U.S. Consulting Services, Inc. www.uscsi.com NonStop Solutions · utimaco · V2verify · Verifone · WORLDLINE · XID TECHNOLOGY SOLUTIONS · XYPRO Mission Critical Security

Thank you for attending this talk TBC23-TB59 How NonStop Excels in the Industry

Roland.lemoine@hpe.com

HPE Slides and Materials Usage

This content is protected

This presentation is the property of Hewlett Packard Enterprise and protected by copyright laws of the United States. The material in this presentation is provided to attendees of the NonStop TBC 2023 as part of their registration and attendance at the event. Attendees are free to use this material and share it with others within their own company.

This material may not be quoted, copied, communicated or shared with third parties or mutual customers without permission from HPE. To request permission to share material in this presentation outside of your company, send an email to roland.lemoine@hpe.com explaining the usage you are intending and your request will be considered.