

ハイパーコンバージドインフラを 利用してみて

2016年11月30日

NTTスマートコネクト株式会社

サービスオペレーション部

炭谷 真也 <sumitani@nttsmc.com>

会社紹介

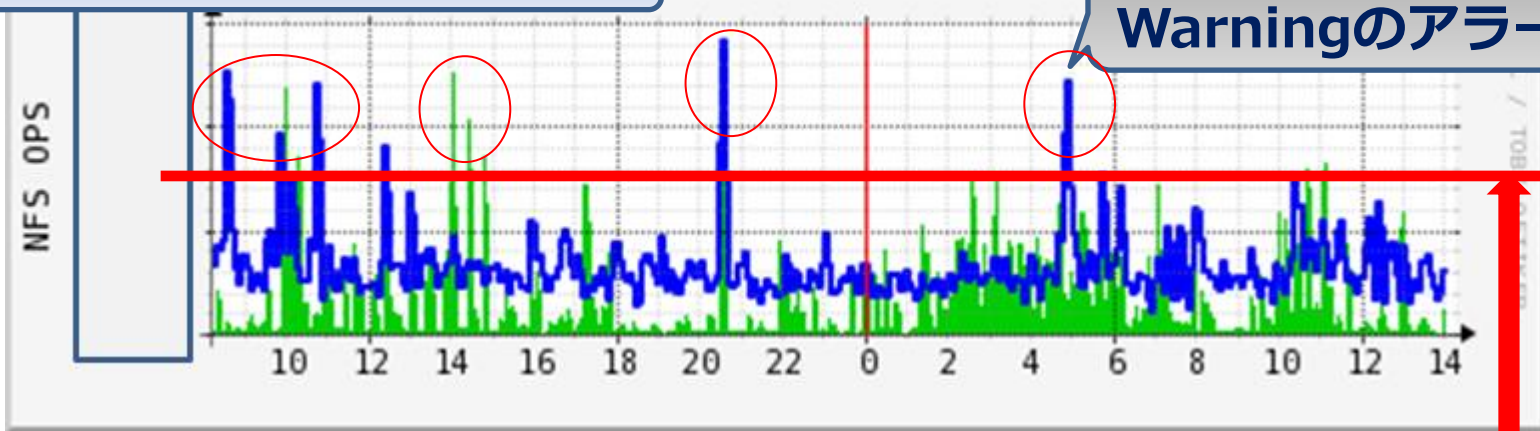
- ◆ 会社名 : エヌ・ティ・ティ・スマートコネクト株式会社
- ◆ 事業内容 : インターネットデータセンター（iDC）を核とした「ハウジング」、「クラウド」、「ストリーミング」サービスの提供
- ◆ 所在地（オフィス） : 大阪市北区大深町3番1号 グランフロント大阪タワーC 13階
- ◆ データセンター拠点 : 大阪（堂島ほか）
東京・名古屋・広島・福岡ほか
- ◆ 設立時期 : 2000年3月
- ◆ 資本 : 1億円（NTT西日本100%）
- ◆ 社員数 : 183名（H28.7現在）
- ◆ 社員平均年齢 : 36.3歳



ハイパーコンバージドインフラ利用の背景

- 2012年12月 vSphereベースのIaaSサービスを提供開始
 - ※開始当初は3層型構成(サーバ/ストレージ/NW)でスタート
- IaaSサービスが軌道に乗り、ある程度VM台数が増えたところでストレージのIOPSがボトルネックとなったインシデントが増加

既存ストレージのIOPSグラフ



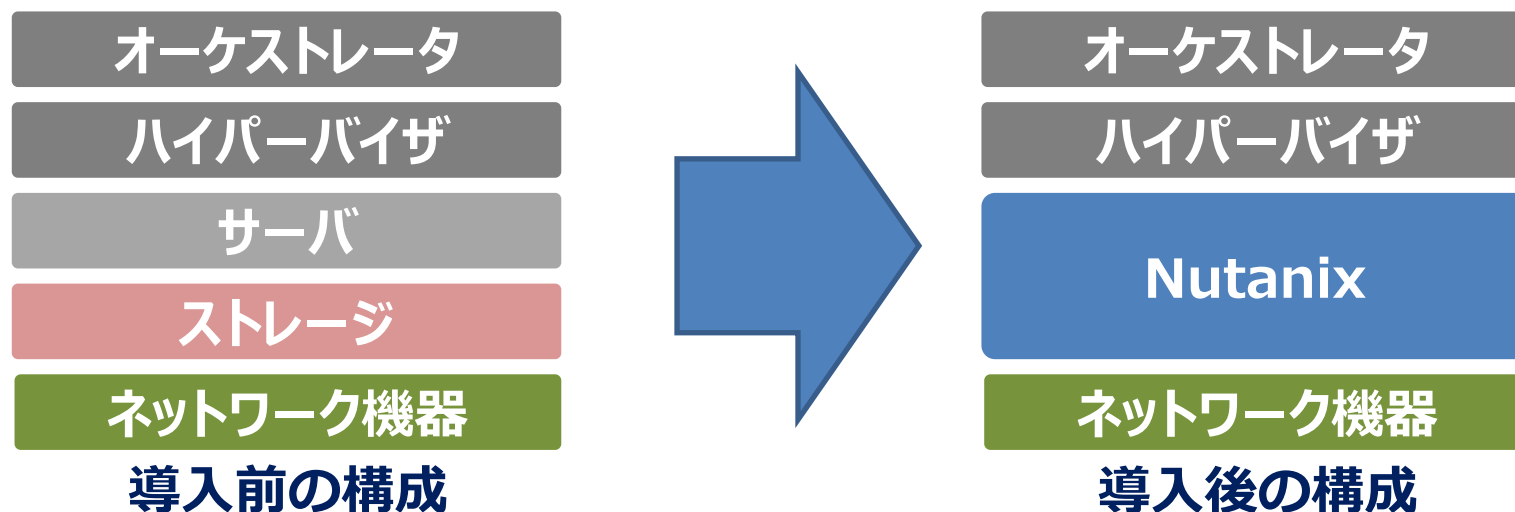
※【縦軸】 IOPS 【横軸】 時間

※既存ストレージのベンダが推奨するIOPS限界値

- インシデント解消を目的としてストレージ、インフラ構成の見直しを検討
- 2014年 機種選定・検証を経て、HCIである**Nutanix**の商用導入を決定

Nutanix選定時のポイント

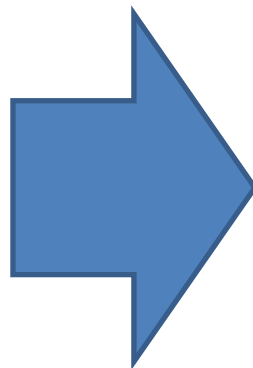
- 3層型の選定候補と比べても、IOPS単価が最安
- 構成がシンプル、スケールアウトの簡単さ
 - 3層型の場合、スケールさせるとどうしてもコストが増大する試算
 - 設計、構築、運用コストの削減
- サービスプロバイダとしての対外的PR
 - 当時は国内サービスプロバイダでの導入実績はなし



➤ ストレージIOPS、ラック収容効率が向上



導入前



比較項目	従来比
ストレージIOPS	2.5~5.0倍
ラック収容効率	2.5倍以上

導入後

- **設計時**

- ストレージの設計コストが従来と比較した場合、ほぼゼロ

- **構築時**

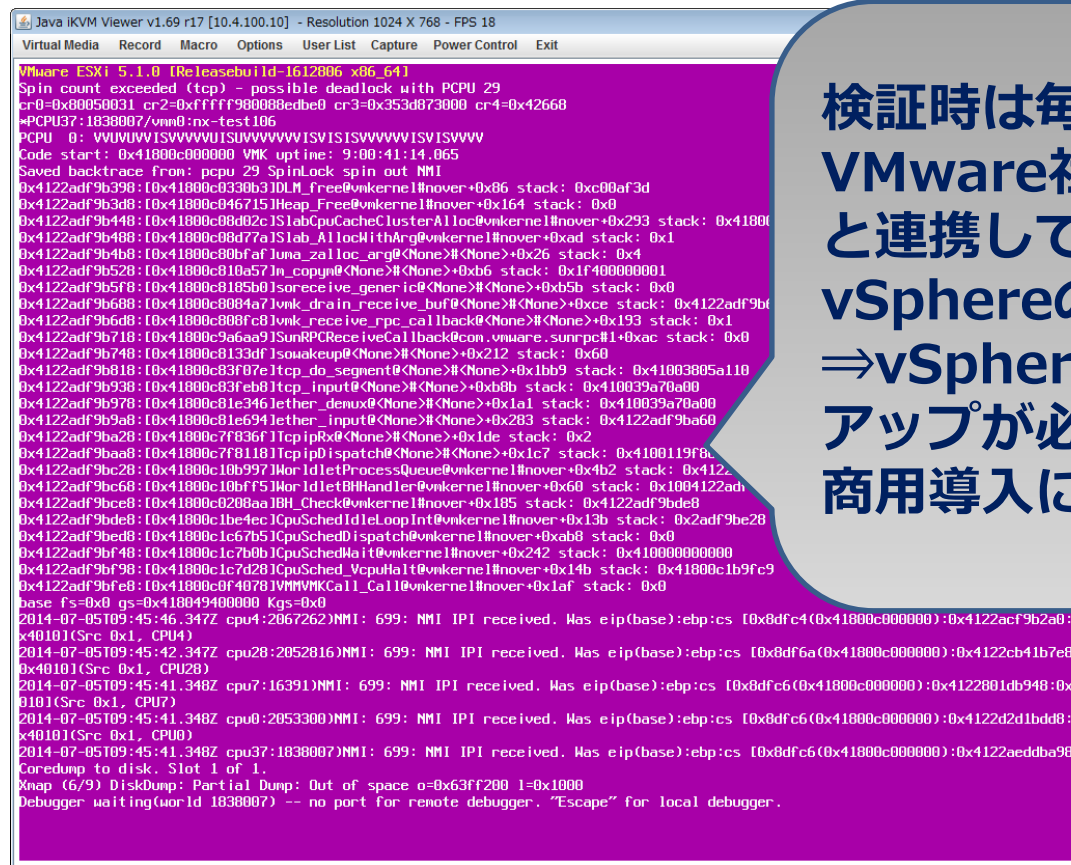
- 必要な作業はラッキング、ケーブリング、初期セットアップに削減
 - ハイパーバイザのインストール、基本的なネットワーク設定はキッティングされ、出荷されてくる
- 初期セットアップ、スケールアウトが簡単なため、構築トータル時間の削減
- ラック収容効率向上によるハウジングコストの削減

- **運用時**

- パフォーマンス起因のインシデント件数減少による稼働削減

● 検証時のトラブルケースの紹介

➤ 導入要件に沿った検証は実施しましょう



```
Java iKVM Viewer v1.69 r17 [10.4.100.10] - Resolution 1024 X 768 - FPS 18
Virtual Media Record Macro Options User List Capture Power Control Exit

VMware ESXi 5.1.0 [Releasebuild-1612006 x86_64]
Spin count exceeded (tcp) - possible deadlock with PCPU 29
cr0=0x80050031 cr2=0xfffff980088edbe0 cr3=0x353d873000 cr4=0x42668
*PCPU37: 1838007/vmm0:nx-test106
PCPU 0: VVVVVVVVSVVVVVVVVSVVVVVVVVSVVSVVVVVVVVSVVSVVVVV
Code start: 0x41800c800000 VMK uptime: 9:00:41:14.065
Saved backtrace from: pcpu 29 SpinLock spin out NMI
0x4122adf9b398: 0x41800c8330b3!IDL_M_free@vmkernel!#nover+0x86 stack: 0xc00af3d
0x4122adf9b3d8: 0x41800c046715!Heap_Free@vmkernel!#nover+0x164 stack: 0x0
0x4122adf9b448: 0x41800c08d02c!SISlabCpuCacheClusterAllloc@vmkernel!#nover+0x293 stack: 0x41800
0x4122adf9b488: 0x41800c08d77a!SISlab_AllocWithArg@vmkernel!#nover+0xad stack: 0x1
0x4122adf9b4b8: 0x41800c80bfa7!luna_zalloc_arg@<None>#<None>+0x26 stack: 0x4
0x4122adf9b520: 0x41800c810a57!ln_copyin@<None>#<None>+0xb6 stack: 0x1f400000001
0x4122adf9b5f8: 0x41800c8185b0!soreceive_generic@<None>#<None>+0xb5b stack: 0x0
0x4122adf9b680: 0x41800c8004a7!vmk_drain_receive_buf@<None>#<None>+0xc stack: 0x4122adf9b
0x4122adf9b6d0: 0x41800c800fc8!vmk_receive_rpc_callback@<None>#<None>+0x193 stack: 0x1
0x4122adf9b718: 0x41800c9a6aa9!SunRPCReceiveCallback@com.vmware.sunrpc#1+0xac stack: 0x0
0x4122adf9b748: 0x41800c8133df!sowakeup@<None>#<None>+0x212 stack: 0x60
0x4122adf9b818: 0x41800c83f07e!tcp_do_segment@<None>#<None>+0x1bb9 stack: 0x41003805a110
0x4122adf9b938: 0x41800c83f0e8!tcp_input@<None>#<None>+0xb8b stack: 0x410039a70a00
0x4122adf9b978: 0x41800c81e346!ether_denux@<None>#<None>+0x1a1 stack: 0x410039a70a00
0x4122adf9ba58: 0x41800c81e694!ether_input@<None>#<None>+0x283 stack: 0x4122adf9ba60
0x4122adf9ba28: 0x41800c7f836f!tcp_rx@<None>#<None>+0x1de stack: 0x2
0x4122adf9baa8: 0x41800c7f8118!tcp_dispatch@<None>#<None>+0x1c7 stack: 0x4100119f8
0x4122adf9bc28: 0x41800c10b997!WorldletProcessQueue@vmkernel!#nover+0x4b2 stack: 0x412
0x4122adf9bc68: 0x41800c10b9f5!WorldletBHHandler@vmkernel!#nover+0x60 stack: 0x1004122ad
0x4122adf9bce8: 0x41800c0208aa!BH_Check@vmkernel!#nover+0x185 stack: 0x4122adf9bde8
0x4122adf9bde8: 0x41800c1be4ec!CpuSchedIdleLoopInt@vmkernel!#nover+0x13b stack: 0x2adf9be28
0x4122adf9bed8: 0x41800c1c67b5!CpuSchedDispatch@vmkernel!#nover+0xab8 stack: 0x0
0x4122adf9bf48: 0x41800c1c7b0b!CpuSchedWait@vmkernel!#nover+0x242 stack: 0x410000000000
0x4122adf9bf98: 0x41800c1c7d28!CpuSchedVcpuHalf@vmkernel!#nover+0x14b stack: 0x41800c1b9fc9
0x4122adf9bfe8: 0x41800c0f4078!VMVMKCall_Call@vmkernel!#nover+0x1af stack: 0x0
base fs=0x0 gs=0x418049400000 Kgs=0x0
2014-07-05T09:45:46.347Z cpu4:2067262)NMI: 699: NMI IPI received. Was eip(base):ebp:cs [0x8dfc4(0x41800c000000):0x4122acf9b2a0:0
x40101(Src 0x1, CPU4)
2014-07-05T09:45:42.347Z cpu28:2052816)NMI: 699: NMI IPI received. Was eip(base):ebp:cs [0x8df6a(0x41800c000000):0x4122cb41b7e8:
0x40101(Src 0x1, CPU28)
2014-07-05T09:45:41.348Z cpu7:16391)NMI: 699: NMI IPI received. Was eip(base):ebp:cs [0x8dfc6(0x41800c000000):0x4122801db948:0x4
0101(Src 0x1, CPU7)
2014-07-05T09:45:41.348Z cpu0:2053300)NMI: 699: NMI IPI received. Was eip(base):ebp:cs [0x8dfc6(0x41800c000000):0x4122d21dbdd8:0
x40101(Src 0x1, CPU0)
2014-07-05T09:45:41.348Z cpu37:1838007)NMI: 699: NMI IPI received. Was eip(base):ebp:cs [0x8dfc6(0x41800c000000):0x4122aeddba98:
Coredump to disk. Slot 1 of 1.
Xnap (6/9) DiskDump: Partial Dump: Out of space o=0x63ff200 1=0x1000
Debugger waiting(worId 1838007) -- no port for remote debugger. "Escape" for local debugger.
```

検証時は毎日PSODが発生
VMware社、Nutanix社
と連携して解析した結果、
vSphereの問題(※)でした
⇒vSphereバージョン
アップが必要だった結果、
商用導入に遅れ

※ Nutanix社のエンジニアBlogにも掲載されています(「vSphere 5.1 Update 1でTCP Heapが枯渇し、PSODとなる」)
<http://longwhiteclouds.com/2014/07/11/heads-up-purple-diagnostic-screen-due-to-tcp-heap-exhaustion-in-vsphere-5-1-update-1/>

- Nutanixの分散ファイルシステムの容量とデータ可用性の設計

- 独自の分散処理により、クラスタ内に単一ストレージプールを提供
- 標準はRF=2 ※実効容量：50% 可用性：N+1

Q.RAID-6の共有ストレージから移行した場合、データ可用性は悪くなる？

表：RF=2構成時のストレージプールの維持可否

故障ケース	ストレージプールの維持可否(○/×)
a ノードが1台故障	○
b ノード内の1ディスクが故障	○
c ノード内のCVM1台がダウン	○
d 1ブロックが故障	× ※1ブロック=1ノードモデルの場合除く
e ノードが2台同時に故障	×
f 異なるノードの1ディスクが同時に故障	×
g 異なるノードでa,b,cの2つが同時に発生	×

- データ可用性は自社要件に照らし合わせて検討が必要

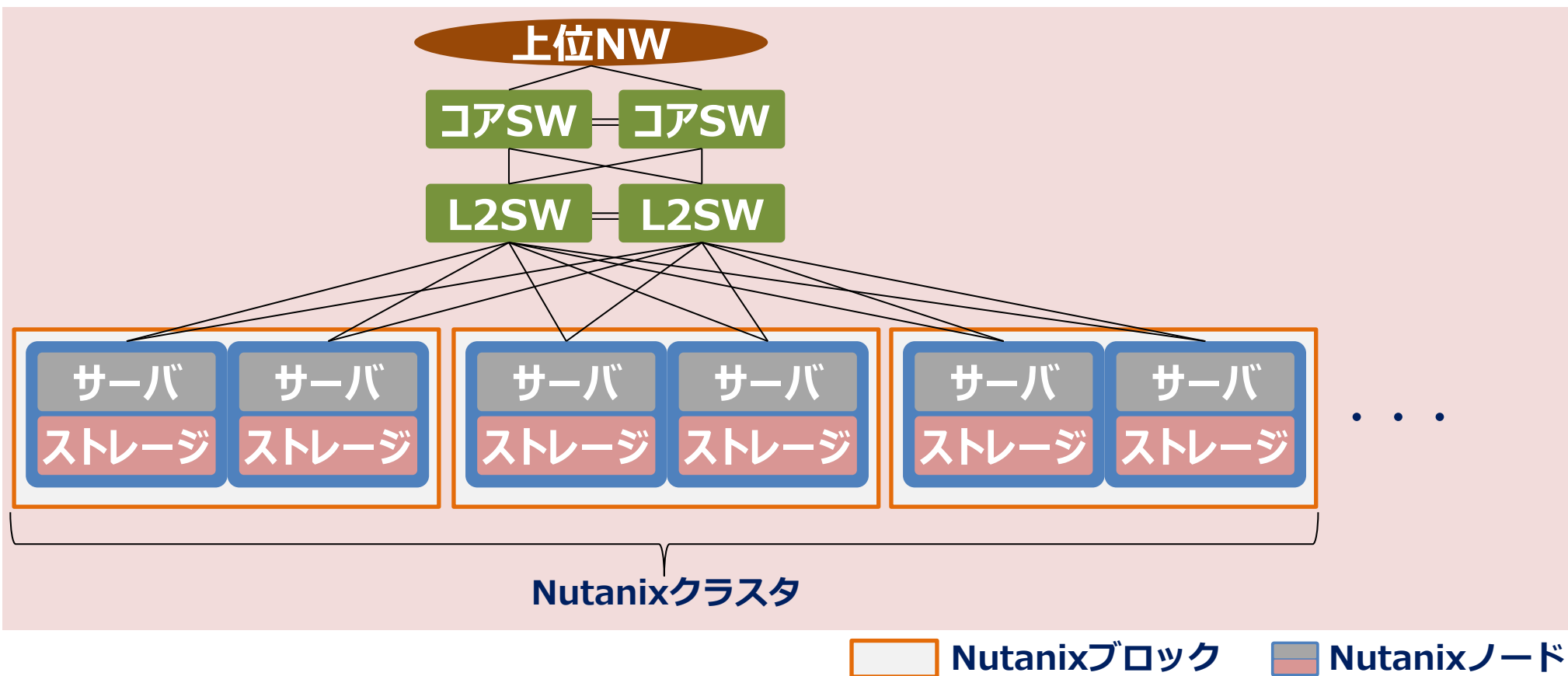
- Block Fault Toleranceを構成する際は1ブロック分の容量を差引いて計算

	RF=2	RF=3	Block Fault Tolerance RF=2	Block Fault Tolerance RF=3
実効容量	◎ 50%	△ 33%	○ 50% – 1ブロック分の容量	× 33% – 1ブロック分の容量
データ可用性	× N+1 ・1ノード ・1ディスク	○ N+2 ・2ノード ・2ディスク	△ N+1(最大N+4) ・1ノード ・1ディスク ・1ブロック	◎ N+2(最大N+4) ・2ノード ・2ディスク ・2ブロック
クラスタ構成時の最小条件(※)	・3ノード以上	・5ノード以上 ・2ブロック以上	・同一ノード数のブロックを3ブロック	・同一ノード数のブロックを5ブロック

※ クラスタ構成条件の詳細は「Product Mixing Restrictions」、 「Data Resiliency Levels」の項参照
http://download.nutanix.com/documentation/v4_7/Hardware_Admin_Ref-AOS_v4_7.pdf

➤ Nutanix構成例

- Block Fault Tolerance + RF=2
- エンタープライズが求めるミッションクリティカル要件を満足させるためにブロック故障が発生してもサービス継続可能な構成



• まとめ

- Nutanixを導入することで、ストレージIOPS、ラック収容効率が向上
- 導入時は自社要件に沿った検証の実施、データ可用性の検討がポイント

• サービスプロバイダとしての今後

- HCIの機能を用いたサービスの提供 ※他社HCI も含めてです
 - DRサービス
 - マルチハイパーバイザの提供検討