

# 超高速データベース専用サーバ

**Insight Dube**

SSDを搭載するだけでは  
速くならない

**Insight Technology**  
no challenge, no life

# Disclaimer

本資料の内容はすべて執筆者の個人的な見解であり、株式会社インサイトテクノロジーの公式な見解を示すものではありません。

本資料内で使用されている社名、ロゴ、製品、サービス名およびブランド名は、該当する各社の登録商標または商標です。

本資料内でベンチマークプログラムの一部を流用して性能測定していますが、データベースのベンチマーク目的ではありません。

本資料の一部あるいは全体について許可無く複製および転載することを禁じます。

# 自己紹介



- 石川 雅也 (いしかわ まさや)
- 現職
  - 株式会社インサイトテクノロジー 取締役 CTO  
FB: MasayaIshikawa
- 職歴
  - 外資系HWベンダーにて、ミニコンOS, UNIX OS担当SE. 社内SE向け技術情報発信、トラブルシュート。crash dump解析、UNIX SCSI Device Driverの改造なども。
  - 1993年 日本オラクルに入社。DBインフラ系テクニカルサポート、コンサルティング、トラブルシュート。
  - 1995年 インサイトテクノロジー設立  
Oracleパフォーマンス管理ツール Performance Insight開発 (1995~)  
Oracleアクセスログ取得管理ツール PISO開発 (2004~)  
SQL Server版、富士通Symfoware版 PISO開発 (2005~)  
DB コンサルティング、トラブルシュート

**2009年 SSDのR&D開始**

**2011年1月 DB向けHW Insight Qube R&D開始。**

**2012年3月 Insight Qube発売**

# 会社概要

- 社名
  - 株式会社インサイトテクノロジー
- 設立
  - 1995年7月
- 資本金
  - 1億円（2013年7月）
- 役員
  - 代表取締役 小幡 一郎
  - 取締役 石川 雅也
  - 取締役 永見 和平
  - 取締役 森田 俊哉
  - 取締役 張 統
- 本社
  - 東京都渋谷区恵比寿1-19-19  
恵比寿ビジネスタワー5F
- 大阪支店
  - 大阪府大阪市北区大深町3番1号  
グランフロント大阪  
ナレッジキャピタルタワーC 11F
- 事業内容（後頁に詳細）
  - コンサルティング事業
  - ソフトウェア事業
  - ハードウェア事業
- 主要取引先
  - 株式会社アシスト
  - 新日鉄住金ソリューションズ株式会社
  - 富士通株式会社
  - 日本電気株式会社
  - 株式会社日立ソリューションズ
  - 伊藤忠テクノソリューションズ株式会社
  - 三井情報株式会社
  - 日本オラクル株式会社
  - マイクロソフト株式会社
  - デル株式会社
  - 日本HP株式会社

# 事業詳細

## コンサルティング事業

対象データベース: Oracle, SQL Server, MySQL など

支援内容: データベースの設計、構築、運用支援、パフォーマンス改善など

## ソフトウェア事業



データベースのセキュリティ保全ツール(自社開発)

6年連続シェアNo1



大規模システムの統合監視ツール(自社開発)



データベースのパフォーマンス監視・分析ツール(自社開発)



マルチデータベースのリアルタイムレプリケーションツール



Oracle Standard Edition向け DR構築ツール



次世代型データウェアハウス向け 超高速データベース

## ハードウェア事業 (2012年～)

Insight Dube

世界最速のデータベース専用高速マシン(自社開発)

# データベーステクノロジーイベントの主催

db tech showcase 2012



db tech showcase 2013  
Tokyo

be a Database Tongalist

～とんがったデータベースエンジニアになれ～

- No Sales Talk
- 年2回開催(東京・大阪)
- 3日間で 50+セッション
- 毎回、延べ3000人以上の技術者が参加
- 11月に秋葉原で第6回目を開催。国内DB関連会社20社以上が協賛

db tech showcase 大阪  
技術者魂が燃えます

世界中のデータベース技術者の  
"Rock"な叫びと生意気さがここに



# HWもわかるDBエンジニアに

HWまで理解している DBエンジニアって、  
重宝されるゾ♪

このセッションでは、DB専用ハードウェア  
Insight Qubeの開発を通じて得た知見を、  
SSD, Infiniband, 大容量Memory という 3つ  
の最新技術と共にご紹介します。

# OS, HWも含めた知識

- OS, HWも含めて設計、チューニングするべし
  - カーネルパラメータチューニングはよく行うが、HWとなるとせいぜいシーケンシャルIOとランダムIOを物理的に異なるデバイスに割り振るとか、データとインデックスを別々に格納するとか、その程度。  
もっとHWの特性を理解することで、より良いシステムを作ることが出来る
- 異なるHWの性能を比較できる知識をつけるべき
  - 特にフラッシュ系ストレージ製品で顕著
  - SSDタイプ、PiCeタイプ、フラッシュ・ストレージなどの性能を客観的に比較し、自分のシステムで、どの構成が必要でコストパフォーマンスが良いのかを正しく判断できるようにする
- 最新HWでは今までと異なるチューニング知識が必要
  - シーケンシャルとランダムでデバイスを分けるのは、HDD時代の知識であり、SSDを使用するのであれば、その知識は役立たない

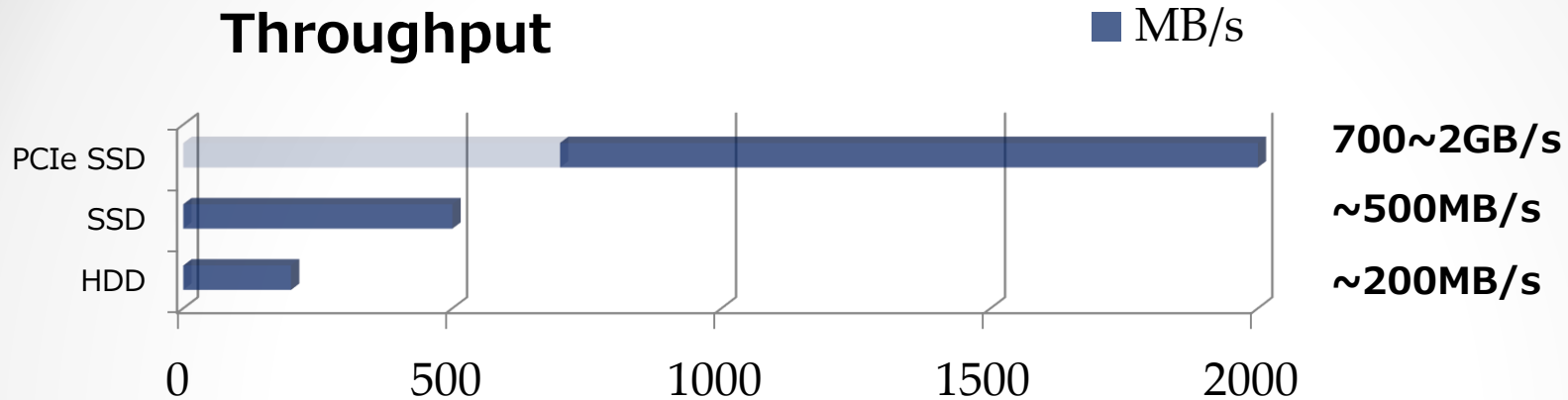


# Agenda of SSD

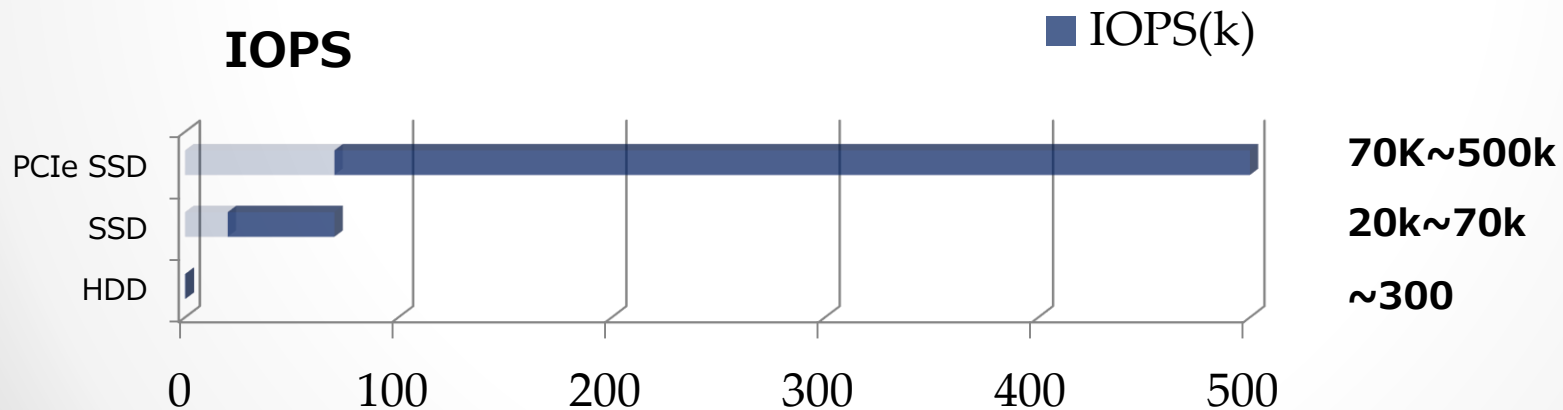
- HDD, SSD, PCIe Flash の比較
- SSDのパフォーマンスを考える
  - 書き込み速度劣化、Client/Enterpriseモデルの違い
  - IOサイズ別パフォーマンスとデータベース
  - カタログスペックに騙されないためには
  - アライメント、OSチューニング、IOスケジューラー など
- SSDの寿命を考える
  - 寿命とは
  - Client/Enterpriseモデルの違い
  - カタログから寿命を読む
- Flashベースの製品を比較する

# Storageの性能比較

## Throughput



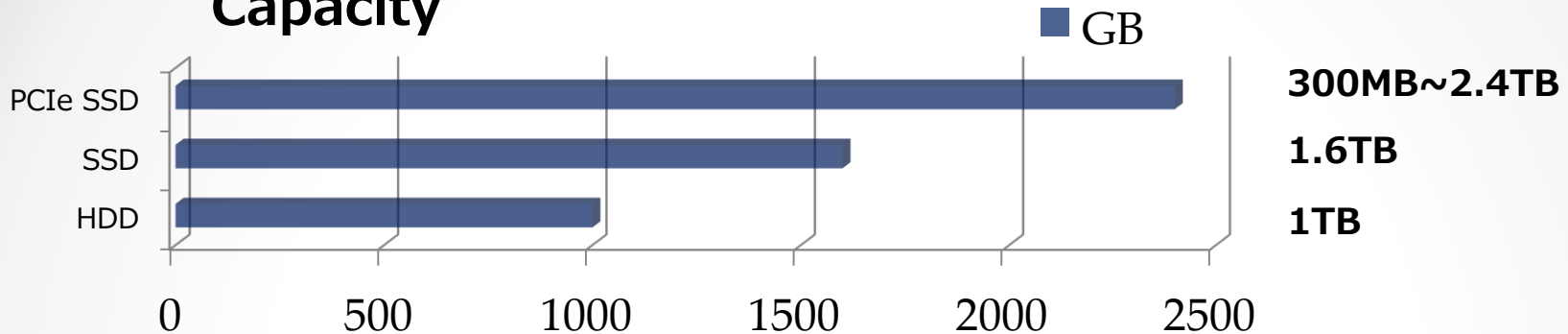
## IOPS



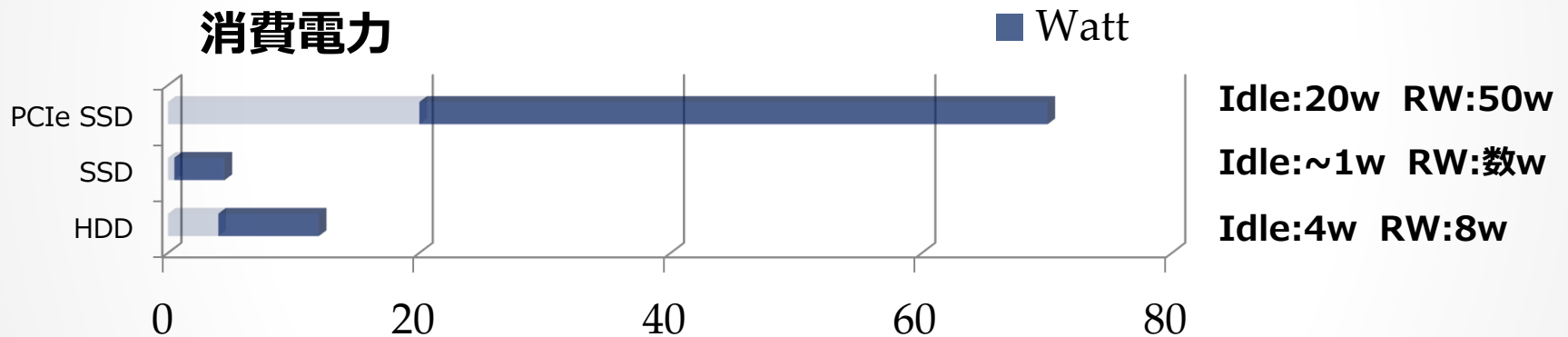
\* コンシューマー向け製品を含む。HDDは 2.5"

# Storageの仕様比較

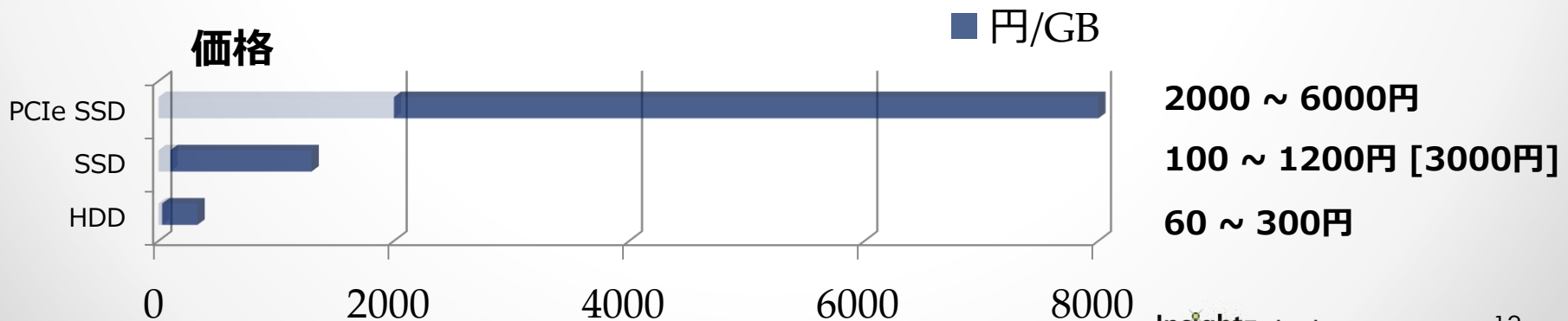
## Capacity



## 消費電力



## 価格

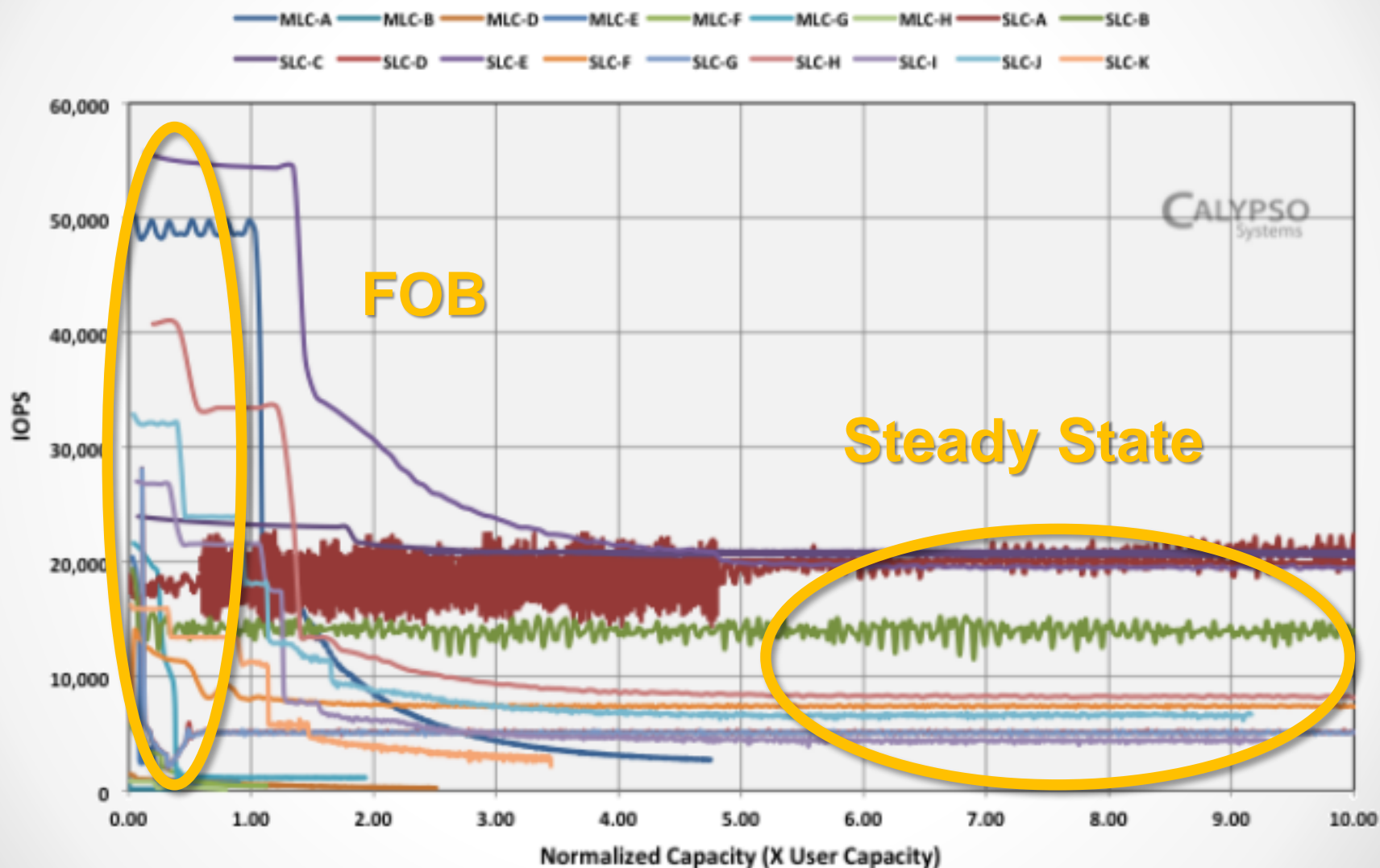


\* コンシューマー向け製品を含む。HDDは 2.5"

# SSDのパフォーマンスを 考える

# SSDのパフォーマンスを考える

## IOPS vs Normalized Capacity



[http://www.snia.org/sites/default/files/UnderstandingSSDPerformance.Jan12.web\\_.pdf](http://www.snia.org/sites/default/files/UnderstandingSSDPerformance.Jan12.web_.pdf)

Copyright © 2013 Insight Technology, Inc. All Rights Reserved.

# SSDのパフォーマンスを考える

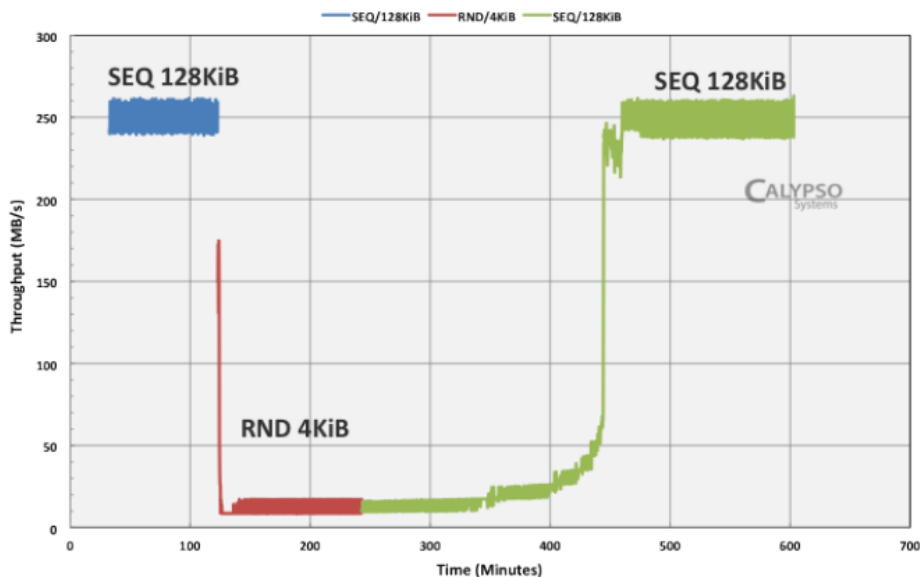
CTS 6.5

Cross Stimulus Recovery: SEQ 128KiB - RND 4KiB- SEQ 128KiB

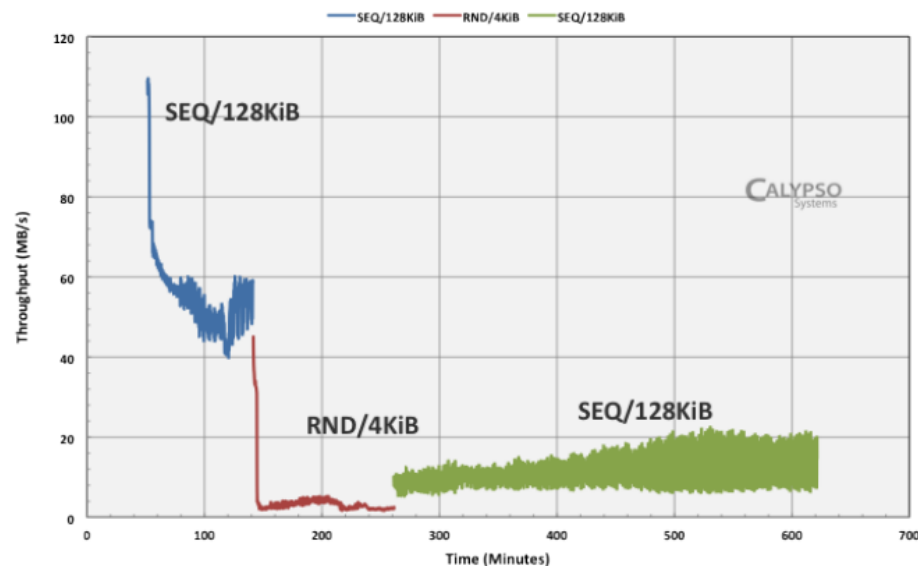
MLC-A - 256 GB

MLC-B - 160 GB

MLC A: XSR (Cross Stim Recovery)



MLC B - XSR: (Cross Stimulus Recovery)



MB/Sec

Measurement Period	Start	After RND 4KiB	End	Measurement Period	Start	After RND 4KiB	End
SEQ 128KiB W	250	5	250	SEQ 128KiB W	120 - 60	10	10 - 20

[http://www.snia.org/sites/default/files/UnderstandingSSDPerformance.Jan12.web\\_.pdf](http://www.snia.org/sites/default/files/UnderstandingSSDPerformance.Jan12.web_.pdf)

Copyright © 2013 Insight Technology, Inc. All Rights Reserved.

# SSDの書き込み速度低下

## Write Amplification

NAND Flash Chipの特性で、データの上書きは出来ず、一度消去してから書き込むしかない。また消去の単位は書き込みの単位より大きい。そのため余計な読み込み・書き込みが発生。

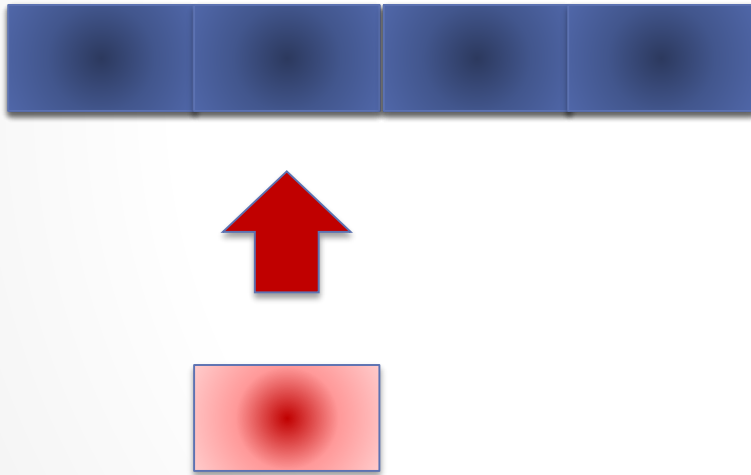


1. データが書き込まれていない場合は、
2. そのまま書き込める

# SSDの書き込み速度低下

## Write Amplification

NAND Flash Chipの特性で、データの上書きは出来ず、一度消去してから書き込むしかない。また消去の単位は書き込みの単位より大きい。そのため余計な読み込み・書き込みが発生。



1. データが書き込まれている場合、
2. 1箇所だけ変更しようとしても、
3. 消去するサイズのデータをコピーし、
4. データを消去し、
5. 1箇所を変更して、



# SSDの書き込み速度低下

## Write Amplification

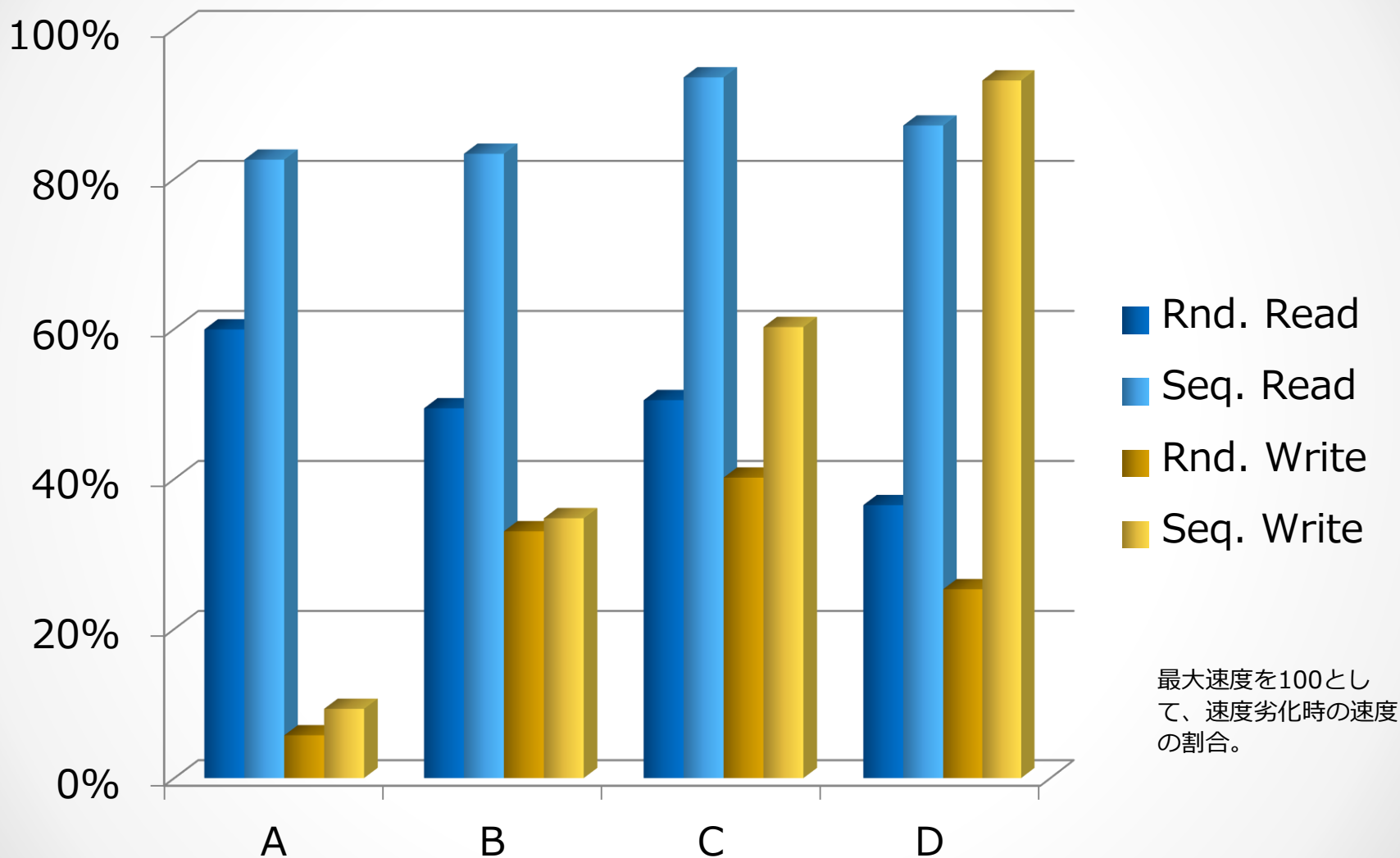
NAND Flash Chipの特性で、データの上書きは出来ず、一度消去してから書き込むしかない。また消去の単位は書き込みの単位より大きい。そのため余計な読み込み・書き込みが発生。



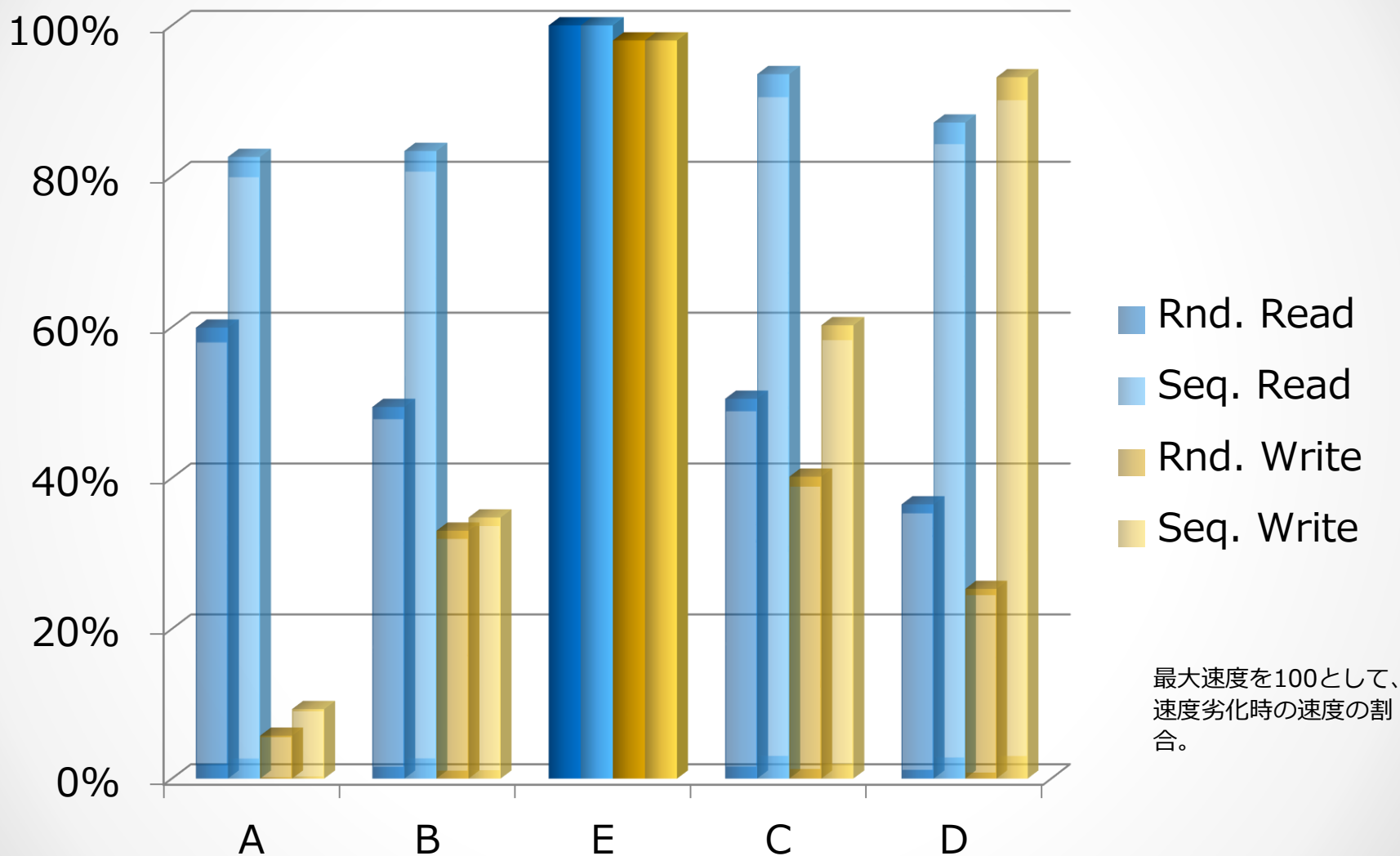
- read: 数10 $\mu$ s
- write: 数10 $\mu$ s ~ 数100 $\mu$ s
- erase: 数ms

1. データが書き込まれている場合、
2. 1箇所だけ変更しようとしても、
3. 消去するサイズのデータをコピーし、
4. データを消去し、
5. 1箇所を変更して、
6. 全て書き戻す。

# SSD製品別 速度劣化率 2011



# SSD製品別 速度劣化率 2012



# DBの典型的なIOサイズ

- Oracle
  - Sequential Read – block size 4KB, 8KB,... の Random Read
  - Scattered Read – block size x multiblock\_read\_count = 128KB, 256KB ... の Sequential Read
  - Direct Path Read – Scattered Readと同じ
- SQL Server
  - Clustered Index Seek – block size 8KB の Random Read
  - Clustered Index Scan – extent size 64KB の Sequential Read
  - 起動オプション設定などの調整で 最大512KB まで可能
- Vectorwise
  - スペック上は 1MB、実測値は 512KB? の Sequential Read
- Filesystem
  - 512B ~

# カタログ・スペックに騙されるな

- Th. R/W 500MB/500MB (大体はIOサイズ128KB~の性能)
- IOPS R/W 2万~7万 (大体はIOサイズ4KB の性能)  
中には 512Bの性能を載せている製品もあるので注意
- 前のチャートで見たとおり製品により特性が異なるので4Kの性能から 8Kの性能は推測できない。
- 自分の使用するDBのIOサイズを見極めて、**そのサイズ  
の性能値**をメーカーに聞くべきである。
- これらの値は、(大体は) FOB(Fresh Out of Box)の性能  
と思え。なので、**Steady (Sustained) Performance**を  
聞け。(心あるメーカーは Steady Performanceを載せて  
いる。)
- だめなら検証するしか無い

# SSD本体以外の考慮点 アライメント



- NAND Flush Memory Blocks
- RAID Stripe Size, Chunk Size
- Partition Alignment
- FileSystem Blocksize
- DB Blocksize

# SSD本体以外の考慮点

## OS関連

- IOスケジューラ (Linux)
  - HDDの場合は cfq
  - SSDの場合は noop (or deadline)
- Kernel options, FileSystem Options, ...

# SSDの寿命を考える



# SSDの寿命を考える

## SSDの寿命とは?

ビットエラーが多発したブロックを不良ブロックとして切り離し、予備領域を使う。予備領域がなくなった時が寿命で、SSDはリードオンリーまたはアクセス不能!! になる。

MLCで3千回から5千回、eMLCで3万回、SLCで10万回が  
NAND Flash Chipとしての書込み寿命(P/E 回数)

MLCは、このままではEnterpriseに使用できないので、  
いかに寿命を延ばすかが、各社の技術!!

Enterprise向けSSDはS.M.A.R.T機能で寿命の予測が可能

# SSDの寿命を延ばすには

## Ware Leveling

書込みを並列化させ、書込み速度の向上を図るとともに、特定のチップに書込みが偏らないように調整する。  
すべてのチップが同時に書込み寿命を迎えるのが理想的

## 予備領域 (Over Provisioning)

一般的なコンシューマー向けSSDの場合 Binary GiB(ギビバイト) ( $1024^3$ ) と Decimal GB ( $1000^3$ )の違いを利用している。  
例えば 256GBの製品の場合 274.88GiB( $256 \times 1024^3$ ) なので、その差 18.8GBが予備領域 (約 7%)

エンタープライズ向けの場合、25%程度を予備領域として確保している製品が多い。製品によっては 40%~200%以上を予備領域としている製品もある。

# SSDの寿命を延ばすには

## ECCエラー訂正

512byte毎に数ビット(1年前は 4~8bit)のエラーが訂正出来る。

Flashメモリのプロセスルールが微細になると、エラー発生率も上がるため、プロセスルールが進化すると ECCの bit数も大きくする必要がある。

ECCで訂正できるビット数を増やすと、その分チップを長く使うことが出来るので、結果として寿命が延びる。

Fusion-ioの ioDriveは 512byteあたり 39bit  
SandForce SF-2500は 512byteあたり 55bit

# 寿命をスペックから読む

- **WA** – Write Amplification

$$WA = \frac{\text{Flashに書き込まれたデータ量}}{\text{Hostから書き込んだデータ量}}$$

通常1.1～10程度。圧縮機能付きコントローラの場合は 0.2 ～ 7.5  
アクセスメソッドやデータの種類などにより異なる。が、コント  
ローラー(**ファームウェア**)の**良し悪しが一番影響**するところ。  
小さいほうが良い

- Sequential Access: WA値が小さい  
Random Access: WA値が大きい
- 圧縮機能付きコントローラの場合、未圧縮データは WA値が小さい  
圧縮済みデータは WA値が大きい

# 寿命をスペックから読む

- **TBW** – Total Byte Written 書き込み可能 総容量

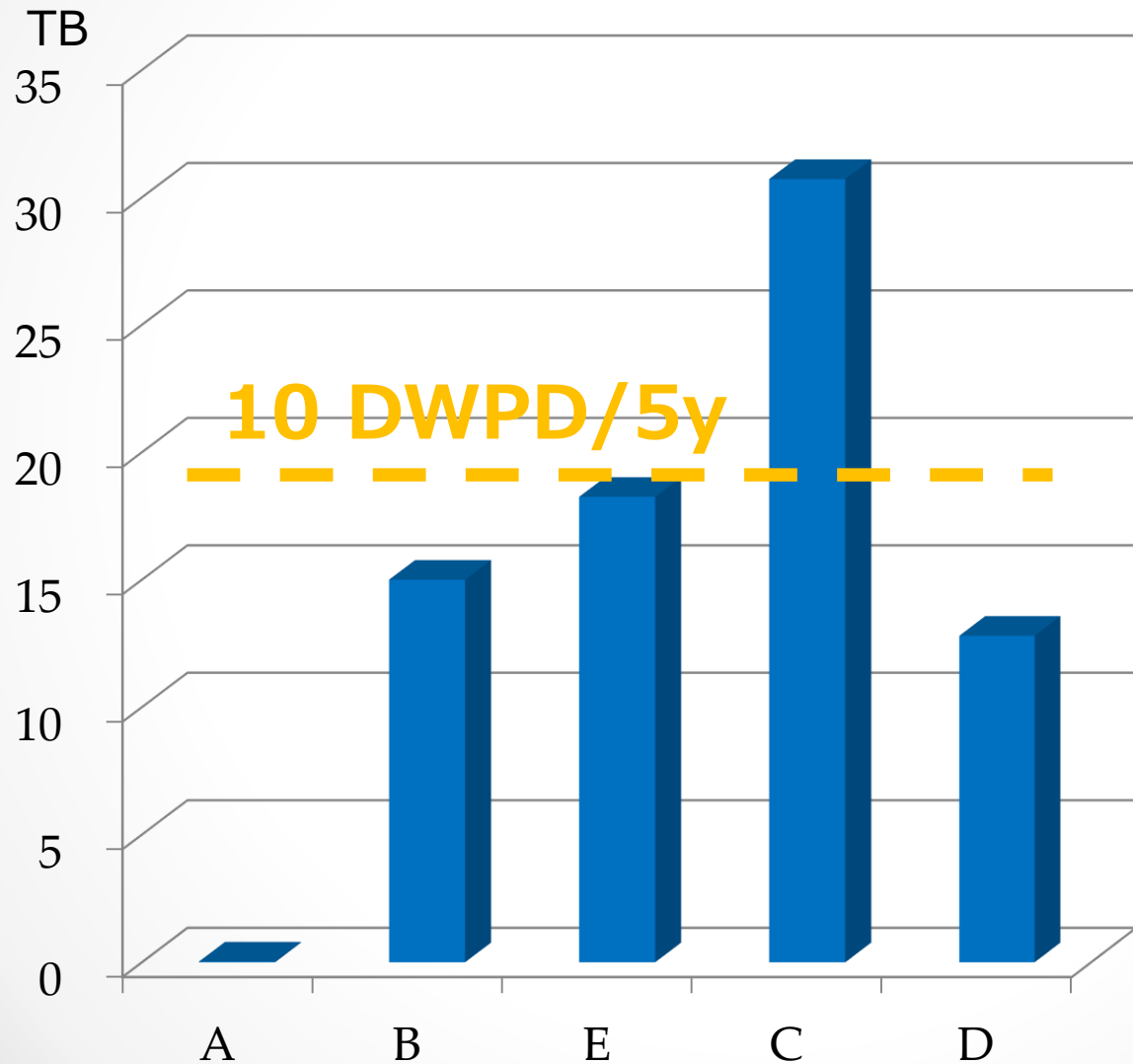
$$TBW = \frac{\text{Capacity(容量)} * (1 + OP) * \text{PE可能回数}}{WA}$$

$$OP = \text{OverProvisioning} = \frac{\text{Physical Capacity}}{\text{Logical Capacity}} - 1$$

- **DWPD** – Data Write per Day  
1日に容量の何倍書き込みして、何年保証するか

$$DWPD = \frac{TBW}{\text{保証期間(日)} * \text{Capacity(容量)}} = \frac{(1 + OP) * \text{PE可能回数}}{WA * \text{保証期間(日)}}$$

# SSD製品別 書き込み寿命



10 DWPD/5y

■ Write Endurance  
TB / GB

容量1GBあたり、何TB書き  
込むと寿命になるか。

# SSD まとめ

- エンタープライズ市場で使えるMLC SSDが多数。
- HDDと比較すると、台数削減が見込めるため、HWコスト、スペースコスト、運用コスト、消費電力コストの削減が実現可能。性能向上による業務へのプラスのインパクトも考えるべき。

# SSD まとめ

- 性能劣化は、ほとんどのSSDで発生する。
- どれ位劣化するかは**カタログには書いていない !!!!**  
劣化の度合いは**値段には比例しない !!!**
- **Steadyなパフォーマンスを教えてください！**
- 寿命もWA値の設定次第でどうにでも見せられる。
- こちらも理論武装して、  
**WA値がいくつで計算した寿命なのか教えてください！**
- 複数の製品を比較できる知識を養おう
- その他、性能を引き出すため、寿命を延ばすために、正しい使い方を。
- 困ったらインサイトテクノロジーに相談 ☺





INSIGHT TECHNOLOGY

EXTREME DATABASE  
PERFORMANCE

# Insight Qube 概要

# Insight Qube とは

## DBパフォーマンス課題

- ◆ 効率の悪いテーブル設計
- ◆ アプリケーションの問題、バインド変数の使い方など
- ◆ 効率の悪いSQL文
- ◆ ロック、ラッチなどのデータベースリソース競合
- ◆ 不適切なデータベースパラメータ
- ◆ システムリソースのボトルネック

InsightTechnology  
no challenge, no life



## コンサルティングで解決

- ◆ 待機イベントの発生回数や待機時間による問題点の特定
  - ◆ 各種データベース統計情報の取得 (v\$表、オブジェクト統計情報)
  - ◆ OSリソースの使用状況調査とボトルネック特定
- ⇒ インデックス作成、データベースパラメータ変更、アプリ改修…

## DBパフォーマンス課題

- ◆ アンバランスなシステムボトルネック(特にDisk I/O)により、本来のCPU性能を發揮できない
- ◆ DBに不利なHWアーキテクチャ
- ◆ データ量増加に伴うパフォーマンスの問題は、単なるチューニングでは解決できない

IO

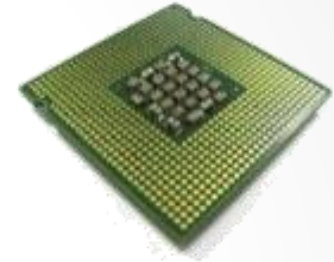


## Insight Qubeで解決

- ◆ Disk I/O「制約」からの解放と大容量メモリによる低レイテンシー処理
- ◆ DB向きなHWアーキテクチャ
- ◆ システムリソースのボトルネックを解消することで大幅な性能向上。
- ◆ DWH専用 RDBMSの提供

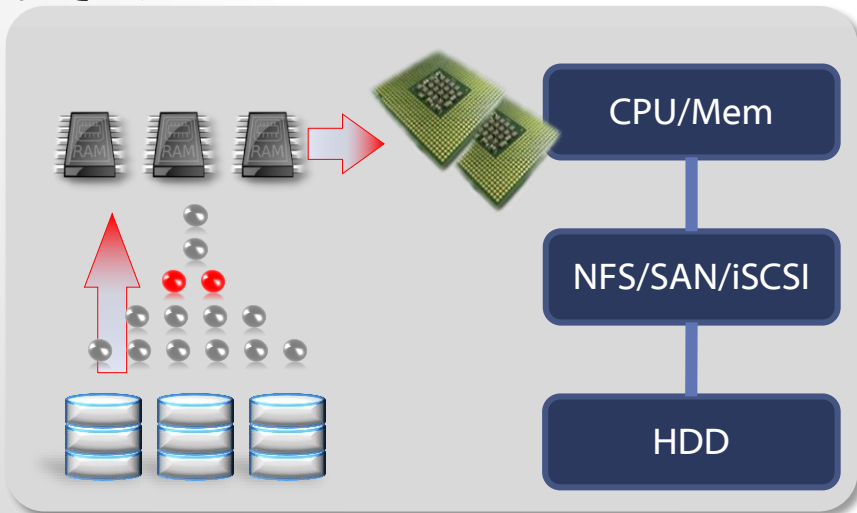
# Insight Qube とは

たった4つしかないデータベースのボトルネック要素(システムリソース)の帯域を拡げ CPU性能を最大限に発揮するよう設計されたサーバーである



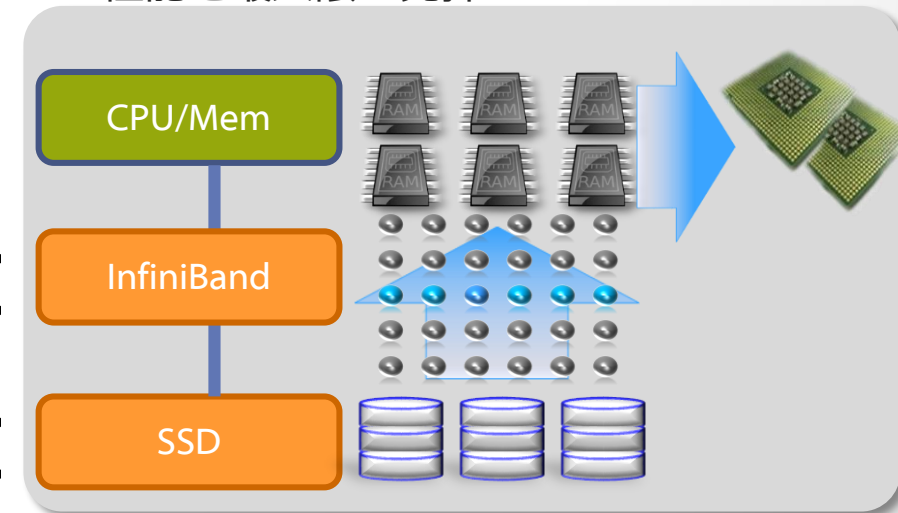
## Low Performance

メモリの容量が少ないためI/Oが多発し、またI/Oの制約によって全体性能は制限されていた



## High Performance

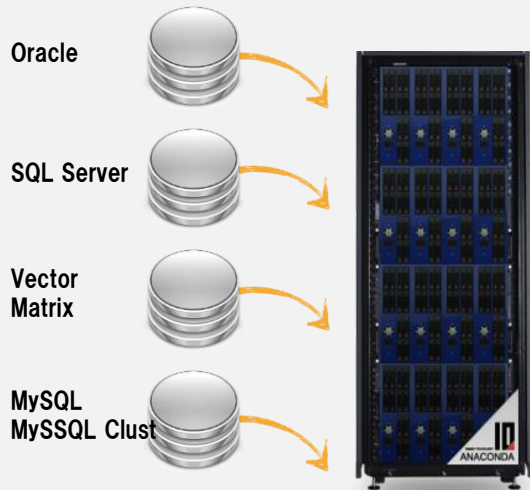
I/O制約を取り除き、さらに大容量メモリによりLatency向上とI/O自体を抑制し、CPU性能を最大限に発揮



# Insight Qubeの特徴

## マルチDB対応

- ◆ Oracle DB (SE One/SE/EE)
- ◆ SQL Server
- ◆ Actian Vector / Matrix (DWH専用RDBMS)
- ◆ MySQL, MySQL Cluster



## データベースに合わせて 2つのタイプを提供 どちらも High Throughput & Low Latency

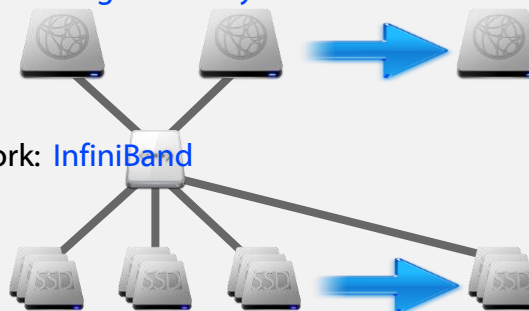
### スケールアウト (クラスター) 構成

- ◆ Oracle RAC, Actian Matrix, MySQL Cluster
- ◆ Oracle最小構成(ストレージノード3台)で**10GB/s**のスループットを実現
- ◆ 2CPUサーバー(DBノード)で最大**768GB**メモリを**リーズナブルな価格**で提供し、**低Latency**を実現。
- ◆ ノードの追加により、**キャパシティとパフォーマンスのスケールアウト**が可能 (ストレージ、DBノード共)

DB node: **Large Memory & Scale Out**

Network: **InfiniBand**

Storage node: **Full SSD & Scale Out**



### スケールアップ (シングル) 構成

- ◆ Oracle, SQL Server, Actian Vector, MySQL, その他
- ◆ 1台で**10GB/s**のスループットを実現
- ◆ 2CPUサーバーで最大**768GB**メモリ 4CPUサーバーの場合 **3TB**メモリを**リーズナブルな価格**で提供し、**低Latency**を実現。

DB node:  
**Large Memory  
Full SSD &  
Scale Up**



# 他社DB専用機と同等性能 ~ OLTP処理

実顧客での PoCテスト内容：

約30GB相当のデータをRAC2ノードからinsert (シーケンシャル/ランダム ライト)のテストを実施

## Insight Qube 構成

<DBサーバ 2台>

Oracle 11gR2, RAC構成 (2node)

CPU: Xeon5690 (3.46GHz/6core)

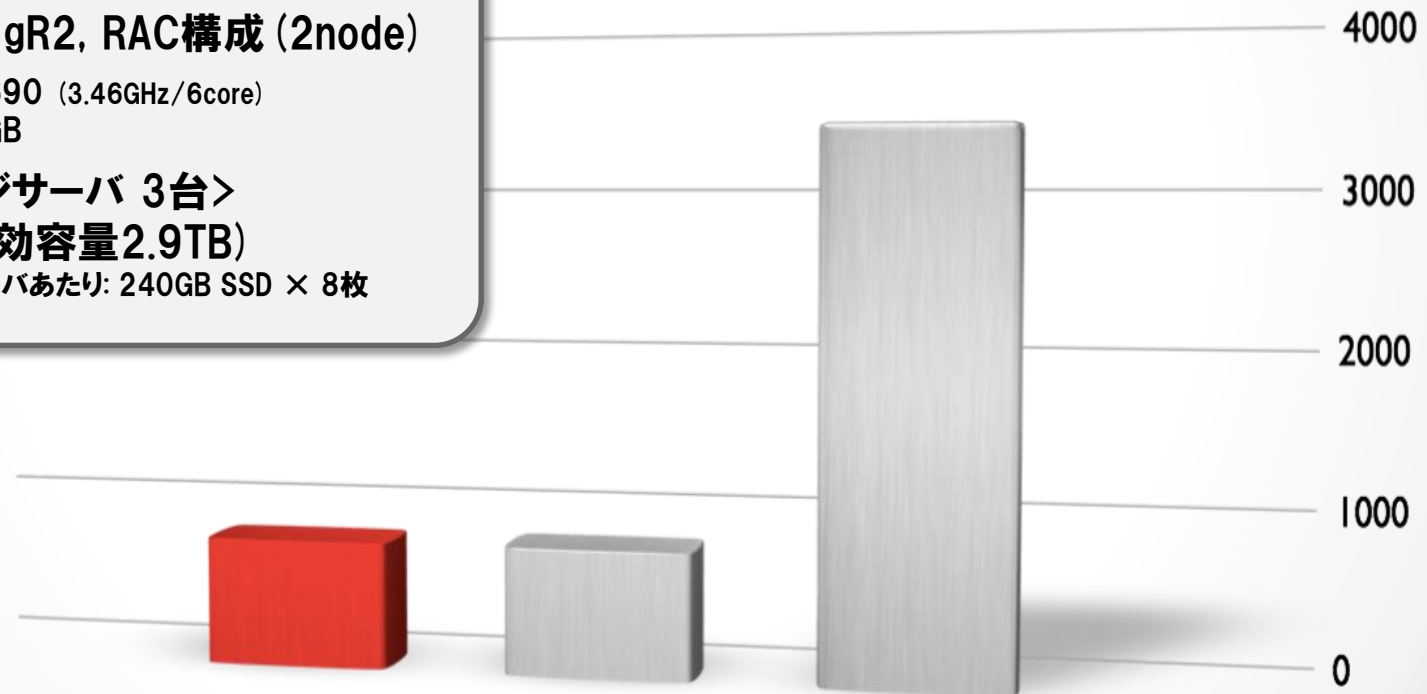
Memory: 48GB

<ストレージサーバ 3台>

5.8TB (実効容量2.9TB)

1ストレージサーバあたり: 240GB SSD × 8枚

処理時間(秒)



Oracle

Insight Qube

DBアプライアンス  
製品

IAサーバー +  
SSDキャッシュ ストレージ

# 他社DB専用機より高性能 ~ Batch処理

実顧客での PoCテスト内容：

元データ 約800GBに対する各種日次/月次バッチ処理の処理時間を比較

処理時間(秒)

## Insight Qube 構成

<DBサーバ 2台>

Oracle 11gR2, RAC構成 (2node)

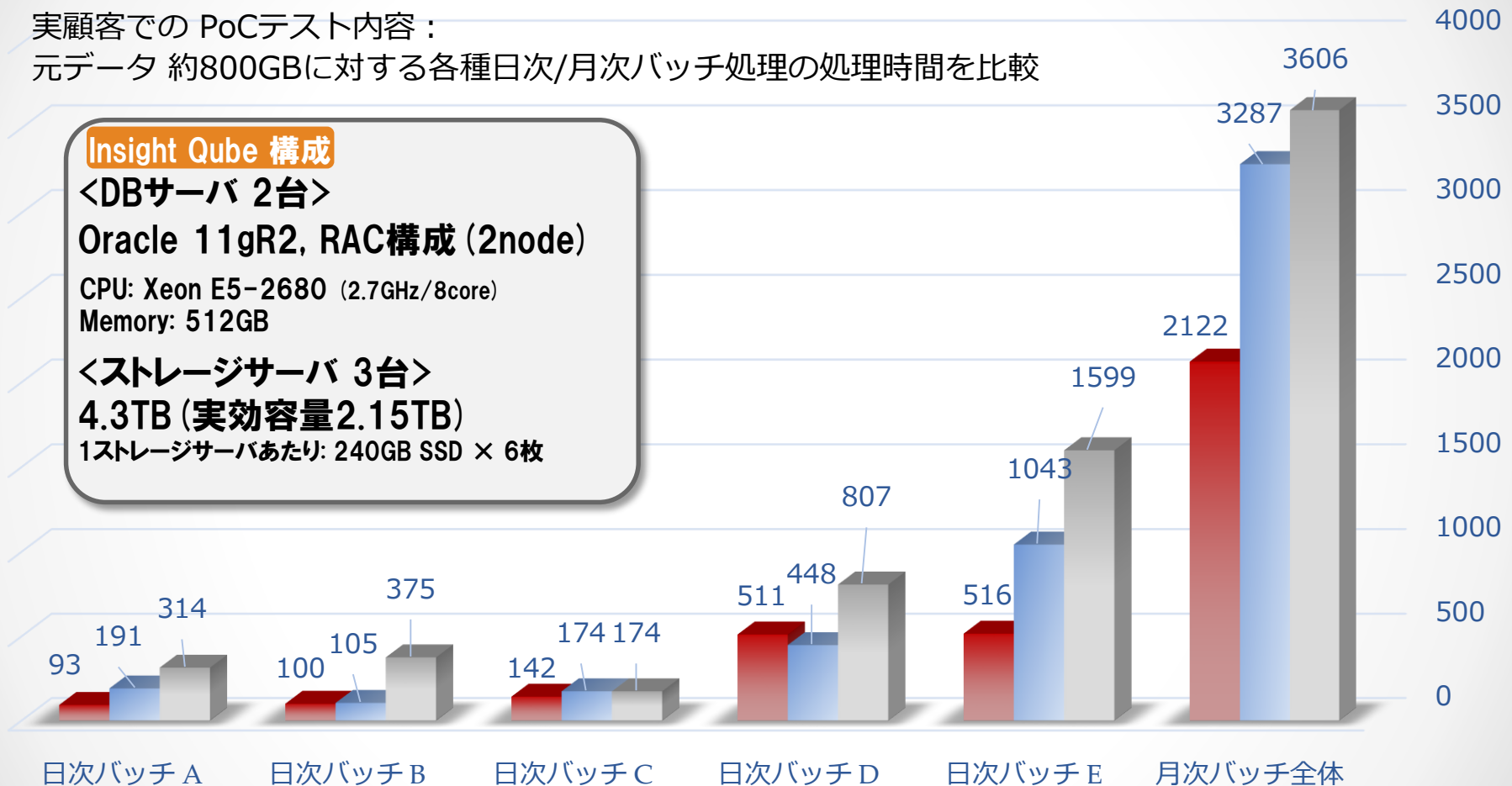
CPU: Xeon E5-2680 (2.7GHz/8core)

Memory: 512GB

<ストレージサーバ 3台>

4.3TB (実効容量2.15TB)

1ストレージサーバあたり: 240GB SSD × 6枚



Oracle

■ Insight Qube

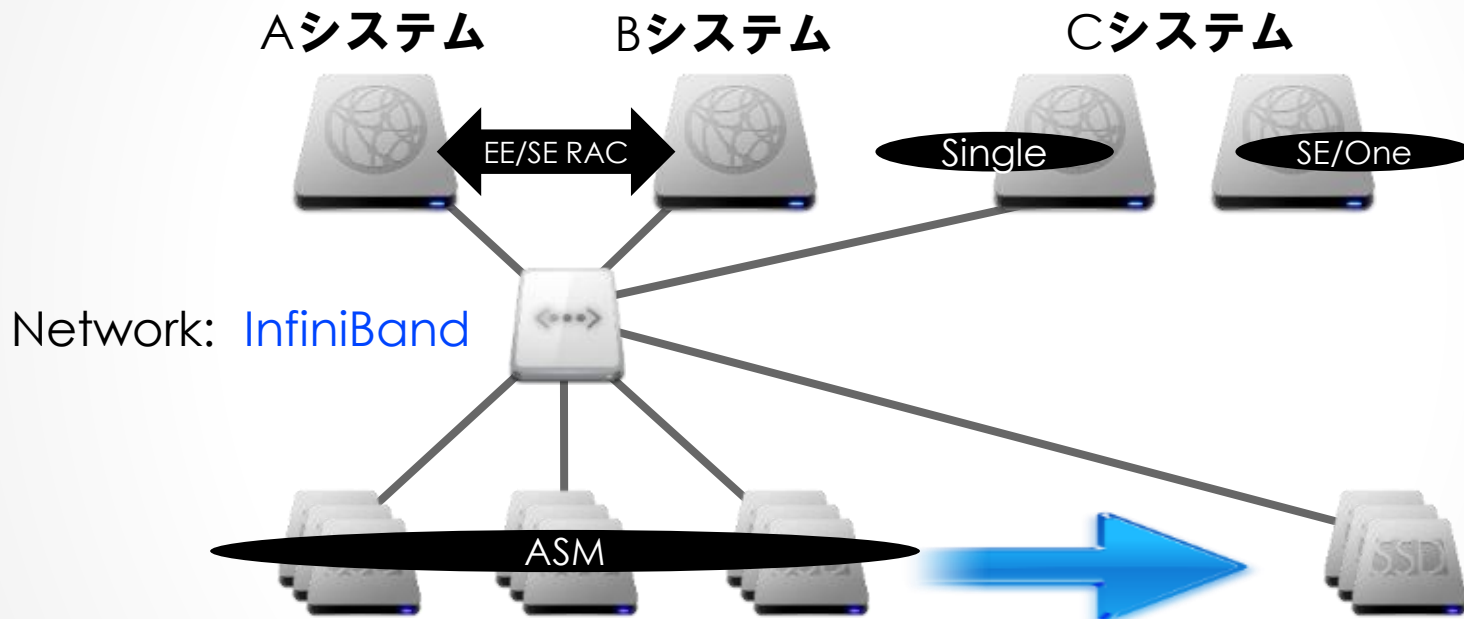
■ DB専用機 (国内メーカー)

■ DBアプライアンス製品

# DB専用機では許されないマルチテナント

DB node: Large Memory & Scale Out

DBノードはマルチテナント  
Oracle RACを構成することも、SE, SE/One も利用可能



Storage node: Full SSD & Scale Out

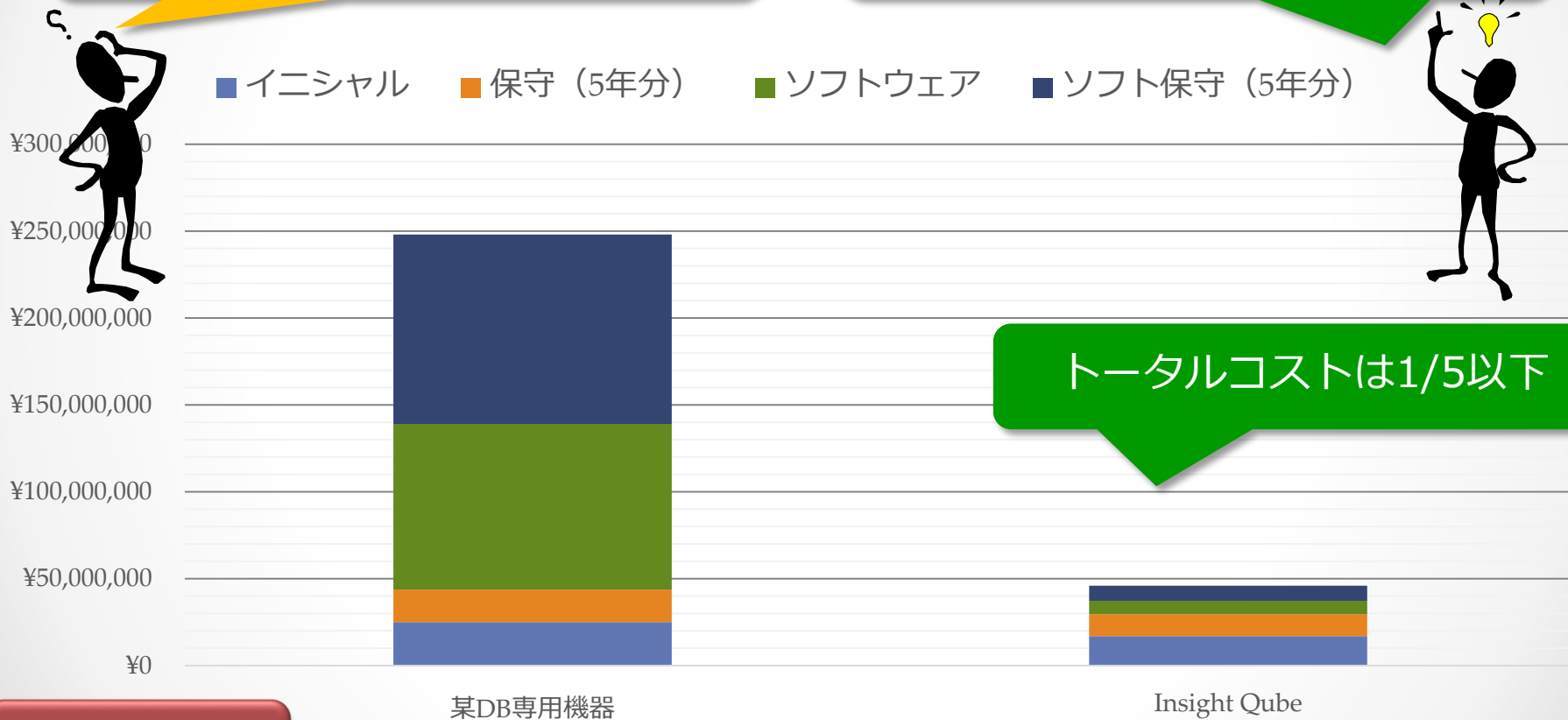
ストレージリソースは共有されるので無駄がない  
ストレージノードを増やすことで全体のパフォーマンスが向上!



# 圧倒的なコストパフォーマンス！

Oracle DB SEを利用してパフォーマンスに課題がある・・・  
某DB専用機器はOracle DB EE・・・

Oracle SEが使える Insight Qubeで、パフォーマンスの課題を解決！




Oracle



# SQL Server®とハードウェアを知り尽くしたプロフェッショナルが作るSSD Appliance Insight Qube® for SQL Server®

高価で高速なフラッシュストレージを使用しても、ファイバーチャネルで接続したアンバランスなシステムではもったいない。SQL Server®を最先端のハードウェアで高速化する2つのポイントを押さえ、OSやSQL Server®の設定をベストチューンすることで、高性能かつコストパフォーマンスに優れたSSD Applianceを実現！



 インテル® Xeon® プロセッサ  
E7-4890v2×4CPU 60コア

■全国どこでも安心の  
24時間365日、最長5年サポート

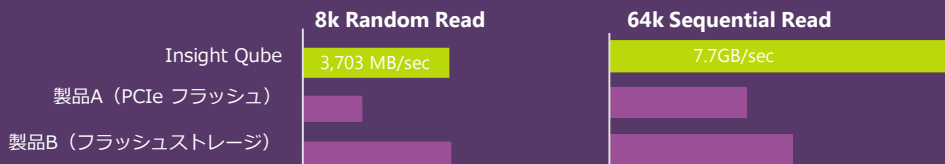
SQL Server は大容量メモリマシンの時代へ。  
コストパフォーマンスならInsight Qube

- SQL Server®2014 で新たに追加されたインメモリOLTPエンジン(コードネーム: Hekaton) を効果的に利用するには大容量メモリが必要です。

2014  
対応

8KBのランダムIO および 64KBのシーケンシャルIOを高速化

- DWH向けの「クラスター化列ストアインデックス」を高速に動作させるために64KBのシーケンシャル/ランダムIOで7.7GB/sのスループットを実現しました。
- 耐久性、コストパフォーマンスに優れたMLC SSDを、**寿命も含め5年間完全保証**。
- 8KBランダムIOも高速化し、OLTP、DWHどちらのシステムでも性能を発揮します。



special  
offer

## エントリーモデル

マシンスペック

- CPU : 32Core (E7-4820v2)
- Memory : 1TB
- Disk:6.4TB (実効3.2TB)

Light OLTP+DWH向け

キャンペーン価格

2014年12月未まで

定価1,533万円

980万円(税抜)

## ハイスペックモデル

マシンスペック

- CPU : 60Core (E7-4890v2)
- Memory : 3TB
- Disk:25.6TB (実効12.8TB)

Heavy OLTP向け

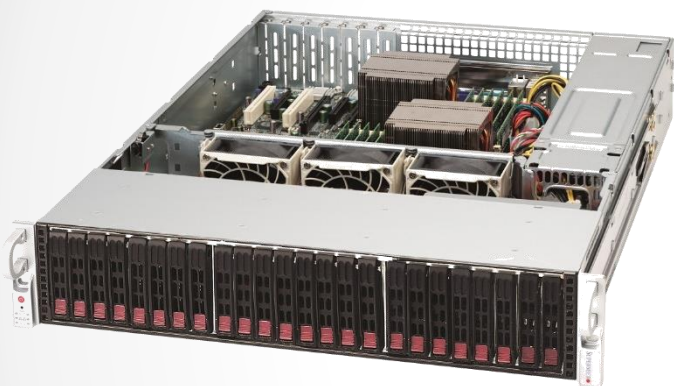
定価 5,613万円(税抜)


詳しくはお問い合わせください。

詳細、お問い合わせはこちらから <http://www.insight-tec.com/products/insight-qube/iq-for-sqlserver.html>

# SQL Server®とハードウェアを知り尽くしたプロフェッショナルが作るSSD Appliance Insight Qube® for SQL Server®

高価で高速なフラッシュストレージを使用しても、ファイバーチャネルで接続したアンバランスなシステムではもったいない。SQL Server®を最先端のハードウェアで高速化する2つのポイントを押さえ、OSやSQL Server®の設定をベストチューンすることで、高性能かつコストパフォーマンスに優れたSSD Applianceを実現！



 インテル® Xeon® プロセッサー  
E5-2697v2×2CPU 24コア

■全国どこでも安心の  
24時間365日、最長5年サポート

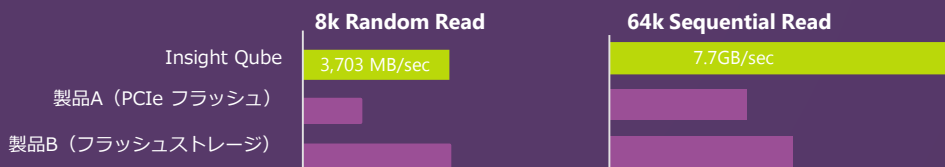
SQL Server は大容量メモリマシンの時代へ。  
コストパフォーマンスならInsight Qube

- SQL Server®2014 で新たに追加されたインメモリOLTPエンジン(コードネーム: Hekaton) を効果的に利用するには大容量メモリが必要です。

2014  
対応

8KBのランダムIO および 64KBのシーケンシャルIOを高速化

- DWH向けの「クラスター化列ストアインデックス」を高速に動作させるために64KBのシーケンシャル/ランダムIOで7.7GB/sのスループットを実現しました。
- 耐久性、コストパフォーマンスに優れたMLC SSDを、**寿命も含め5年間完全保証**。
- 8KBランダムIOも高速化し、OLTP、DWHどちらのシステムでも性能を発揮します。



special  
offer

## エントリーモデル

マシンスペック

- CPU : 24Core (E5-2697v2)
- Memory : 512MB
- Disk:6.4TB (実効3.2TB)

Light OLTP+DWH向け

キャンペーン価格

2014年12月未まで

定価1,133万円

790万円(税抜)

## ハイスペックモデル

マシンスペック

- CPU : 24Core (E5-2697v2)
- Memory : 768GB
- Disk:8.8TB (実効4.4TB)

Heavy OLTP向け

定価 1,880万円(税抜)

詳しくはお問い合わせください。

詳細、お問い合わせはこちらから <http://www.insight-tec.com/products/insight-qube/iq-for-sqlserver.html>

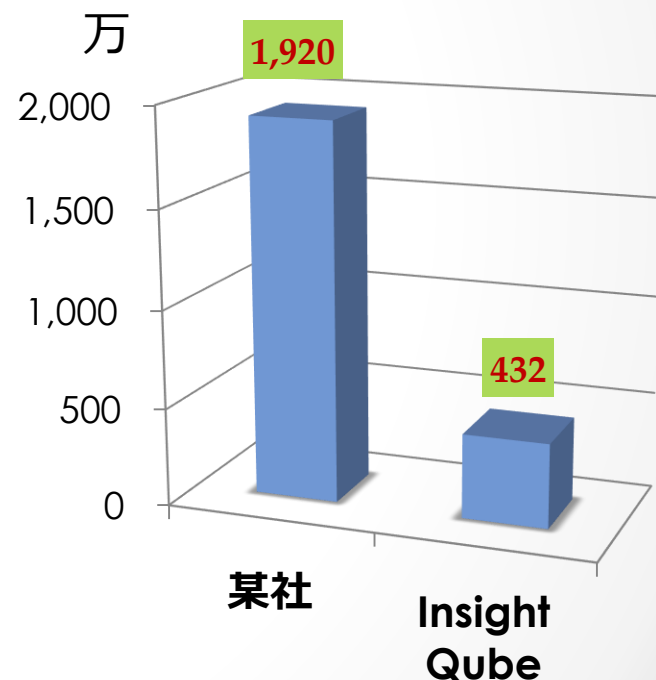
# 大容量メモリでの価格比較

- 大容量メモリを搭載するには32GB LR-DIMMが必要。

例えば Xeon Sandy Bridge  
で最大の768GBメモリを搭載  
するには 32GB x **24枚** 必要

- 某ベンダーと価格を比較すると、Insight Qubeは **1/4**の  
コストで導入可能  
(同じコストでは5枚 160GBしか買えません…)

## 32GB LR-DIMM 24枚の価格





# Insight Technology

no challenge, no life

