30年後の「5分間ルール」

石川　佳治

今回の内容

- 以下の記事を紹介
  - R. Appuswamy, G. Graefe, R. Borovica-Gajic, and A. Ailamaki, The Five-Minute Rule 30 Years Later and Its Impact on the Storage Hierarchy, Communications of ACM, 62(11), 2019.
- 北川先生の教科書（データベースシステム 改訂2版）の演習問題で説明

著者紹介

- Goetz Graefe
  - Microsoft ⇒ HP ⇒ Google
  - DBシステム（オプティマイザなど）

- Anastasia Ailamaki
  - U. Wisconsin ⇒ CMU ⇒ EPFL
  - ハードウェア＋DBなど

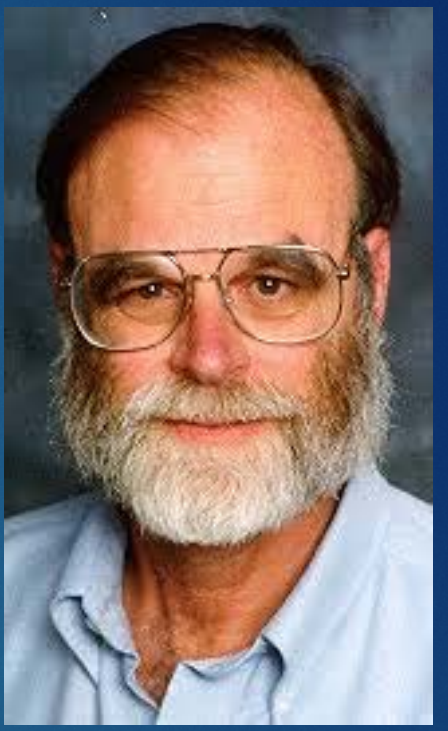
- 両者ともSIGMOD Edger F. Codd Innovation Award受賞

目次

- ジム・グレイについて
- 5分間ルールとは
- ルールの導出
- ルールの適用
- Performance Tierに対する分析
- Capacity Tierに対する分析
- 結論

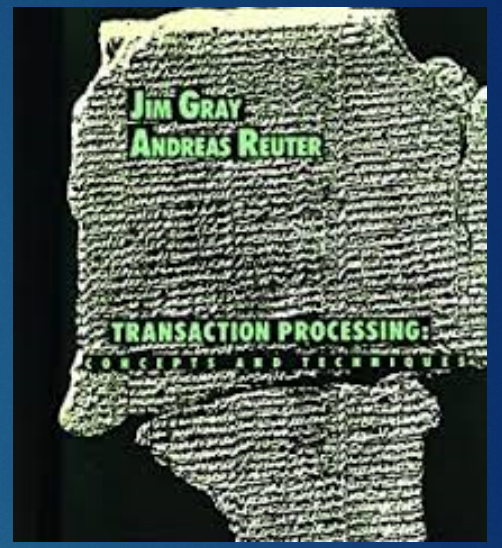
ジム・グレイ（Jim Gray）

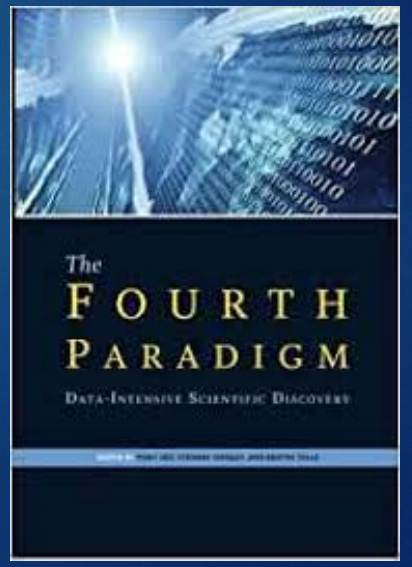
- 1944年～2007年
  - 趣味のヨットで行方不明に
- 経歴
  - IBMでSystem Rの開発などに関わる
  - 1995年よりマイクロソフト
    - 名前をとったMicrosoft Jim Gray Systems Labがウィスコンシン州にある
- チューリング賞（1998年）



ジム・グレイの業績

- トランザクション管理
  - 階層的ロック
  - コミットの意味論
- DWHにおけるデータキューブの概念
- Sloan Digital Sky Surveyのシステム構築
  - 巨大な天文学データベースの構築
- 科学の「第4のパラダイム」の提唱
  - Data-intensive Computing





目次

- ジム・グレイについて
- 5分間ルールとは
- ルールの導出
- ルールの適用
- Performance Tierに対する分析
- Capacity Tierに対する分析
- 結論

５分間ルールの動機

- 大昔の話（1980年代）
- 記憶階層
  - メモリ（DRAM）
    - 高速だが高価，揮発性
  - 磁気ディスク（HDD）
    - 低速だが大容量

- 疑問：経済的な観点からみて，DRAMとHDDをどのように使うのがよいのか？

5分間ルール（1987年）

- 1KBのデータが5分間に1回程度参照される場合，DRAM上に配置すべき
  - つまり，DRAMの容量を適切に増やすなどして，上の条件が成り立つようにすべき
- 当時のハードウェアを前提として算出
- 出典：J. Gray and G. Putzolu, The 5-Minute Rule for Trading Memory for Disc Accesses and the 10-byte Rule for Trading Memory for CPU Time, SIGMOD 1987.

目次

- ジム・グレイについて
- 5分間ルールとは
- ルールの導出
- ルールの適用
- Performance Tierに対する分析
- Capacity Tierに対する分析
- 結論

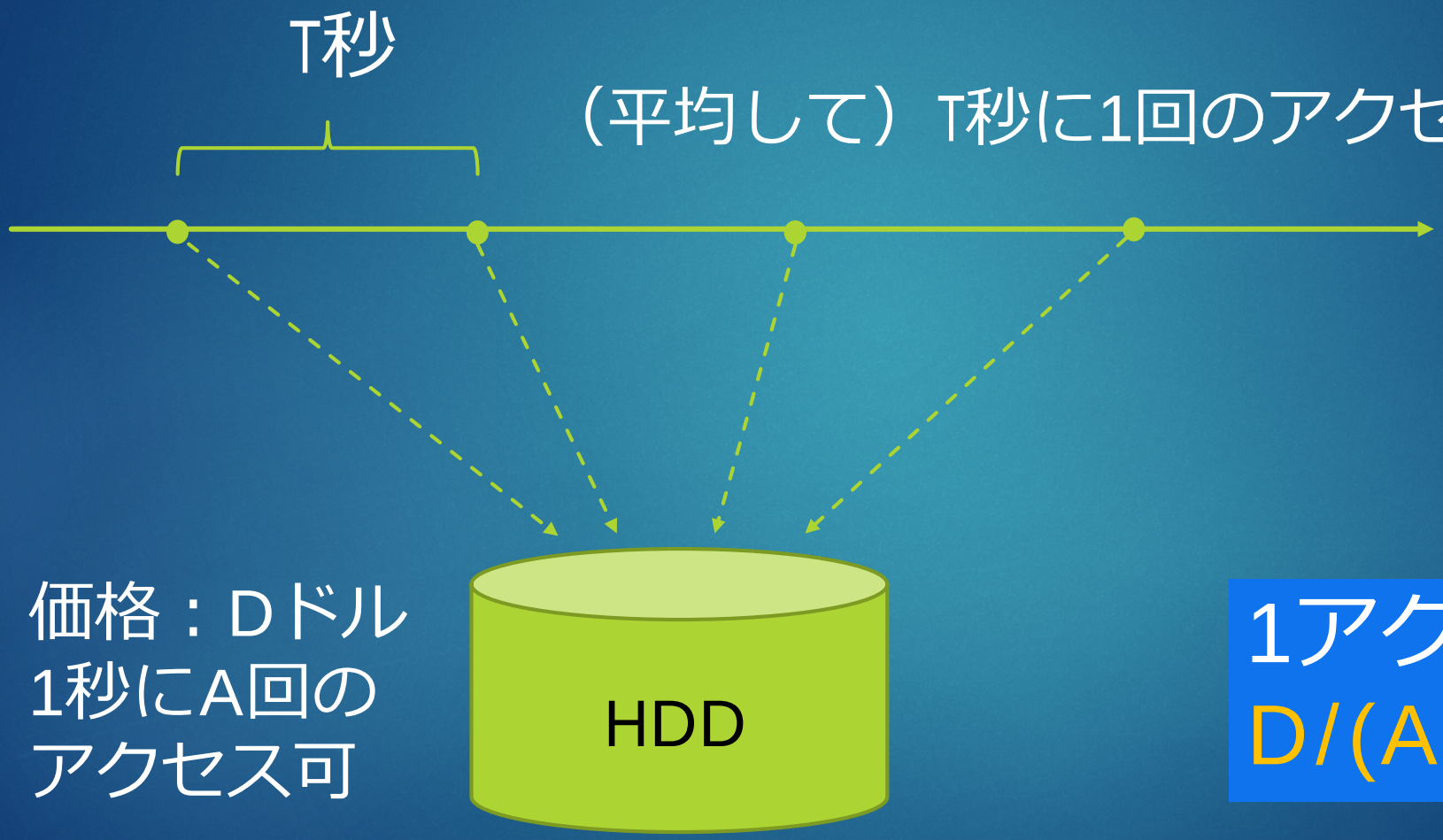
ルールの導出 (1)

- 設定（その1）
  - D：典型的なHDDの価格（単位：ドル）
  - A：1秒間に何回ディスクページへのランダムアクセスを処理できるか（単位：アクセス/秒）
- 1ドル払えば，毎秒A/D（アクセス/秒・ドル）のアクセスが処理可能
- 注：HDDの容量は考慮しない
  - 十分に容量がある，典型的なHDDを想定

ルールの導出 (2)

- 1ドル払えば，毎秒A/D（アクセス/秒・ドル）のアクセスが処理可能 ⇒ 1アクセスにD/A（秒・ドル/アクセス）が必要
- あるページをT秒に1回アクセスするなら（すなわち毎秒1/T回のアクセス），1アクセスあたりD/(A·T)ドルが必要
- ポイント：アクセスに関するコストをお金に換算

ルールの導出：ここまで

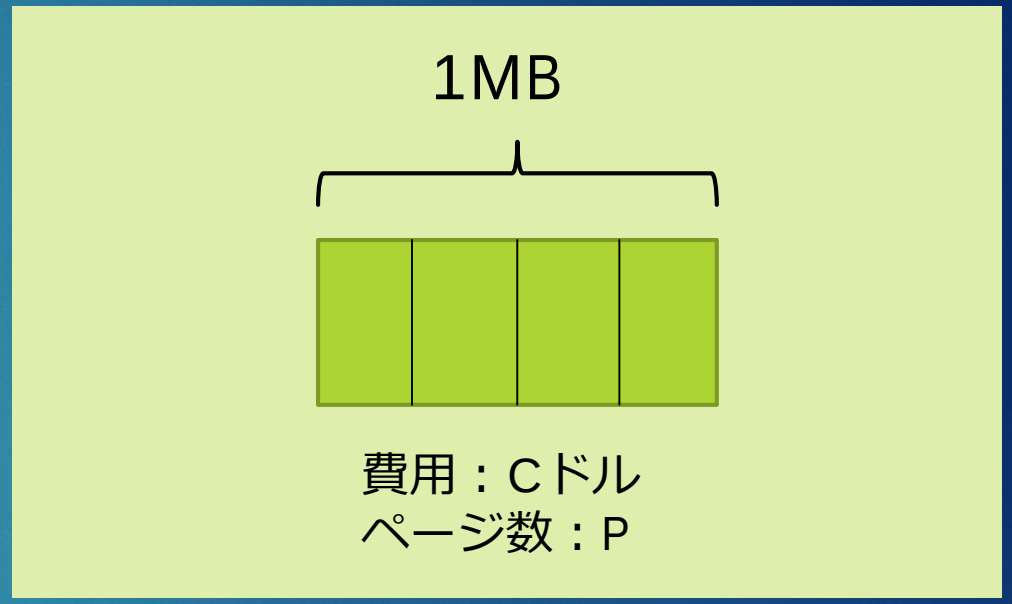


1アクセスあたり
D/(A·T)ドルが必要

ルールの導出 (3)

- 設定（その2）
  - P：1MBのDRAMに入るページ数（単位：ページ）
  - C：1MBあたりのDRAMの価格（単位：ドル）
- DRAMに1ページを確保するためのコストはC/P（ドル）となる

DRAM（主記憶）



つまり，1アクセスを高速に行うためDRAMを用いるコストはC/Pドル

ルールの導出 (4)

- 以下の状況を考える

$$D/(A \cdot T) = C/P$$

  - 左辺：HDDへの1アクセスのために必要な費用
  - 右辺：1アクセスの効率化にDRAMを用いる費用
- 費用の観点からはちょうど釣り合っている

ルールの導出 (5)

- 式を変形

$$T = D \cdot P / A \cdot C$$

- アクセス回数がT秒あたり1回より多いアクセスが発生するようなデータについては，HDDではなくDRAM上に置く（ようにシステムを構成する）方が費用の面で有利

目次

- ジム・グレイについて
- 5分間ルールとは
- ルールの導出
- ルールの適用
- Performance Tierに対する分析
- Capacity Tierに対する分析
- 結論

計算の例：SIGMOD 1987論文の場合

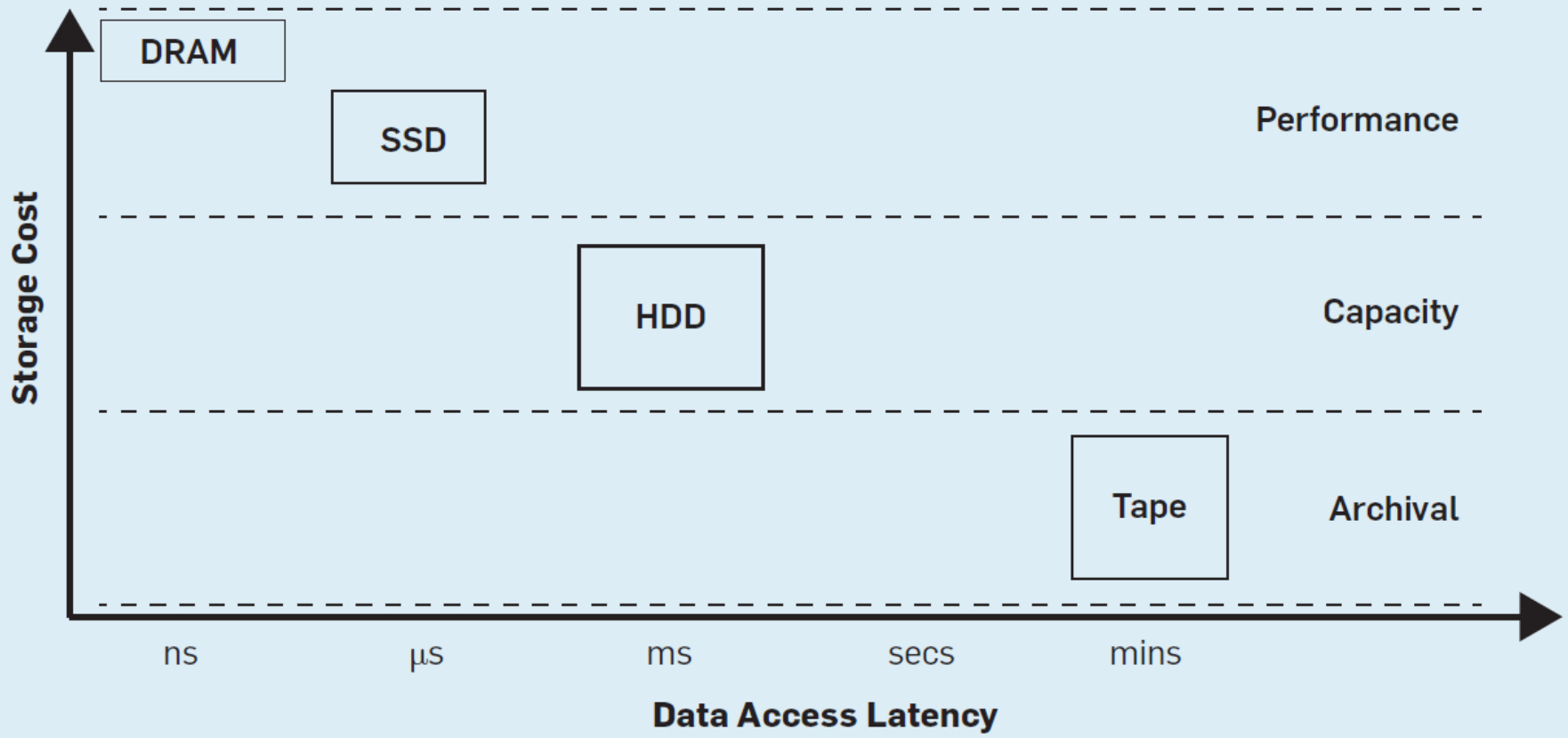
- 設定
  - ページサイズを1KBとする
    - P = 1,024 KB / 4KB = 256
  - HDDの価格 D = 30,000ドル
  - 1秒間のアクセス可能回数 A = 5回
  - 1MBあたりのDRAMの価格 C = 5,000ドル
- T = D·P/A·C = (30,000・256)/(5・5,000)
  = 307.2（秒）= 5.12（分） ← 5分間ルールが得られた！

30年後：5分間ルールはどうなったか

- 30年でハードウェア技術が劇的に変化
- 不揮発性メモリ (non-volatile memory, NVM)
  - フラッシュメモリは携帯機器などで普及
  - SSD (Solid State Drive) も一般化
  - 新たなNVM技術：MRAM（磁気抵抗メモリ）など
- 磁気テープ
  - 大容量のデータを安価に蓄積
  - コールドストレージ（ランダムアクセス不可）

ストレージの階層

Figure 1. Storage tiering for enterprise databases.



余談：期待の不揮発性メモリ技術

- ReRAM（抵抗変化型メモリ），PCM（相変化メモリ），STT-RAM（スピン注入メモリ）

Memory	DRAM	ReRAM	PCM	STT-RAM
Volatile	×	✓	✓	✓
Endurance	10^{15}	10^{8}-10^{11}	10^{8}-10^{9}	10^{12}-10^{15}
Read Latency (ns)	∼10	∼10	20-60	2-35
Write Latency (ns)	∼10	∼50	20-150	3-50
Cell Size (F^2)	60-100	4-10	4-12	6-50
Write Energy(J/bit)	10^{-14}	10^{-13}	10^{-11}	10^{-13}

J. Boukhobza et al., Emerging NVM: A Survey on Architectural Integration and Research Challenges, *Trans. Design Autom. Electr. Syst.*, 23(2), pp. 14:1-14:32, 2018.

余談：Amazon Glacier

- 低コストのオンラインストレージサービス
  - コールドストレージの活用
- Amazon S3 Glacier
  - 1GBあたり月額0.005ドル
  - データの取り出しに，申し込んでから3～5時間
- Amazon S3 Glacier Deep Archive
  - 1GBあたり月額0.002ドル
  - 取り出しに12時間程度

ハードウェアのスペック

- 赤字は1987年のDRAM-HDDの場合に用いた値

Table 1. The evolution of DRAM, HDD, and Flash SSD properties.

C

D

Metric	DRAM				HDD				SATAFlash SSD	
	1987	1997	2007	2018	1987	1997	2007	2018	2007	2018
Unit price($)	5k	15k	48	80	30k	2k	80	49	1k	415
Unit capacity	1MB	1GB	1GB	16GB	180MB	9GB	250GB	2TB	32GB	800GB
$/MB	5k	14.6	0.05	0.005	83.33	0.22	0.0003	0.00002	0.03	0.0005
Random IOPS	–	–	–	–	5	64	83	200	6.2k	67k (r)/20k (w)
Sequential b/w (MB/s)	–	–	–	–	1	10	300	200	66	500 (r)/460 (w)

A

- SSDとしては，SATA接続のフラッシュSSDを想定

計算結果

- 上段：ページサイズを4KBに固定したときのTの値
- 下段：T = 5分間に固定したときのページサイズ

Table 2. The evolution of the page size for which the five-minute rule holds across four decades based on appropriate price, performance, and page size values.

	1987	1997	2007	2018
Break-even (4KB page)	100s	9m	1.5h	4h
Page size (5-minute interval)	1KB	8KB	64KB	512KB

- 2018年ではT = 4h：大抵のデータはDRAMに置くことに

DRAM-SSDの場合

- 同様に計算可能（上段）

Table 3. The evolution of the break-even interval across four decades based on appropriate price, performance, and page size values.

Tier	1987	1997	2007	2018
DRAM–SSD	—	—	15m	5m
SSD–HDD	—	—	2.25h	1.5d

今日では，
5分間ルールは
DRAM-SSDに
ついて成立

- 理由
  - DRAMの低価格化
  - SSDによるランダムIOPSの向上

SSD-HDDの場合

- 下段：SSDをHDDのキャッシュとして利用

Table 3. The evolution of the break-even interval across four decades based on appropriate price, performance, and page size values.

Tier	1987	1997	2007	2018
DRAM–SSD	—	—	15m	5m
SSD–HDD	—	—	2.25h	1.5d

T = 1.5日

何が言えるか

- HDDのSSDへの交換は，費用のみならずデータのキャッシングのためのDRAMの量の削減に効果的
  - DRAM-HDDとDRAM-SSDの場合のTの値（2018年では4.5時間 vs 5分）を比べると，前者では多くのデータをDRAMに持つことになる
- すべてのアクティブなデータはDRAMとSSDで保持するのがよい
  - SSD-HDDの場合 T = 1.5日なので，アクティブでないデータのみHDDで管理

目次

- ジム・グレイについて
- 5分間ルールとは
- ルールの導出
- ルールの適用
- Performance Tierに対する分析
- Capacity Tierに対する分析
- 結論

Performance Tierの状況

- NANDフラッシュ
  - 2000年代半ばでSATAフラッシュSSDが地位を確立
  - 2000年代後半からPCIeフラッシュSSDが出現
    - SATAに対し高速
- NVDIMM
  - DIMM（メモリモジュール）にDRAMだけでなくNANDフラッシュを載せる
- NVM
  - NANDフラッシュより高速である3D XPoint（PCM方式）が製品化

ルールの適用

PCIe接続のNANDフラッシュSSD

Table 4. Price/performance metrics for the NAND-based Intel 750 PCIe SSD and 3D-XPoint-based Intel Optane P4800X PCIe SSD.

Device	Capacity	Price($)	IOPS(k)	B/w(GB/s)
Intel 750	800GB	589	460	2.5
Intel P4800X	480GB	617	550	2.5

3D XPointに基づくPCIe SSD

- ページサイズが4KBの場合，どちらについても
  T = 1分間程度となる

何が言えるか

- NANDフラッシュはさらに価格が低下することが予測されている ⇒ T がさらに小さくなる
- NVMによるSSDの性能向上により T がさらに減少
- SSDはDRAMより消費電力が小さい
  - ここでは考慮していない運用コストの面でも有利
- 以上より，NVMベースのデータベースエンジンへのシフトは避けられない

目次

- ジム・グレイについて
- 5分間ルールとは
- ルールの導出
- ルールの適用
- Performance Tierに対する分析
- Capacity Tierに対する分析
- 結論

HDD

- 一時期はクライダーの法則（Kryder’s law）に従い，ムーアの法則よりも早いペースで容量が増加
  - 13か月でHDDの容量は2倍
- しかし，最近では改善の速度が低下
- 待機時にも電力を消費
  - 企業等では80%のデータはコールドで，その割合は増加している

テープ (1)

- 容量は順調に増大を続ける
- LTO (Liner Tape-Open) という規格：最新はLTO-8

Table 5. Price/performance characteristics of tape.

	1997	2018
Tape library cost ($)	10,000	11,000
Number of drives	1	4
Number of slots	14	10
Max capacity per tape	35GB	15TB
Transfer rate per drive (MB/s)	5	750
Access latency	30s	65s

テープドライブで
数百PBまで運用

テープ (2)

- テープはHDDに比べバンド幅が高い：シーケンシャルアクセスには有利

Table 6. Price/performance metrics of DRAM, HDD, and tape.

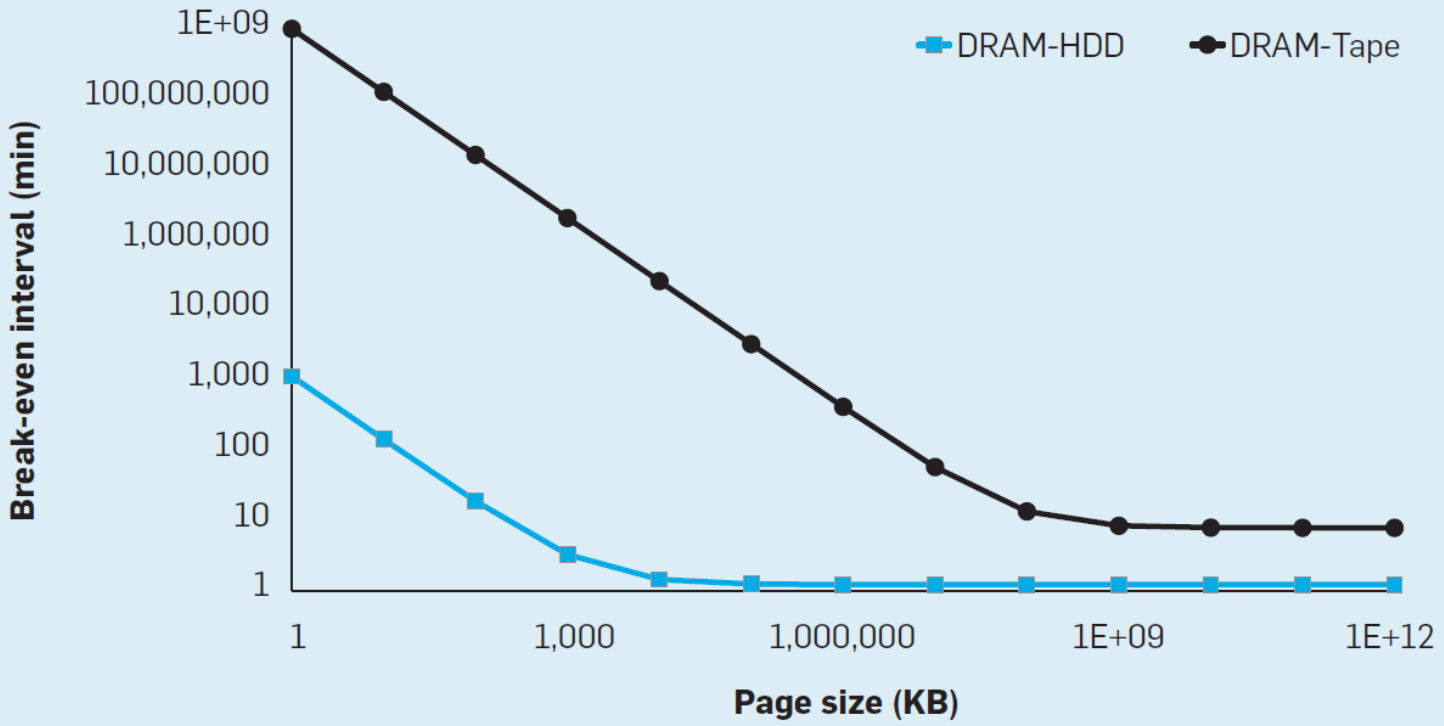
Metric	DRAM	HDD	Tape
Unit capacity	16GB	2TB	10 × 15TB
Unit cost ($)	80	50	11,000
Latency	100ns	5ms	65s
Bandwidth	100 GB/s	200 MB/s	4 × 750MB/s
Kaps	9,000,000	200	0.02
Maps	10,000	100	0.02
Scan time	0.16s	3hours	14hours
$/Kaps	9e-14	5e-09	8e-03
$/Maps	9e-12	8e-09	8e-03
$/Tbscan	8e-06	0.003	0.03
$/TBscan (97)	0.32	4.23	296

一方で、テープのランダムアクセスの遅延はHDDの1000倍

ルールを用いた分析 (1)

- DRAM-HDDとDRAM-tapeについて，ページサイズを変えTを計算
- 4KBのページならDRAM-tapeのTは300年!
  - テープは，チェックインのみのデータモーテル (Gray)

Figure 2. Break-even interval asymptotes for DRAM–HDD and DRAM–tape cases.



ルールを用いた分析 (2)

- DRAM-HDDはT = 1分間に収束（ページサイズ100MB）
- DRAM-tapeはT = 10分間に収束（ページサイズ100GB）
- 転送サイズを極端に大きくしないとペイしない

新たな分析の指標

- MB-access-per-second (Maps)とTBのスキャン (TBscan) がより重要

Table 6. Price/performance metrics of DRAM, HDD, and tape.

Metric	DRAM	HDD	Tape
Unit capacity	16GB	2TB	10 × 15TB
Unit cost ($)	80	50	11,000
Latency	100ns	5ms	65s
Bandwidth	100 GB/s	200 MB/s	4 × 750MB/s
Kaps	9,000,000	200	0.02
Maps	10,000	100	0.02
Scan time	0.16s	3hours	14hours
$/Kaps	9e-14	5e-09	8e-03
$/Maps	9e-12	8e-09	8e-03
$/Tbscan	8e-06	0.003	0.03
$/TBscan (97)	0.32	4.23	296

Mapsに着目：
- DRAMが最も安価
- DRAMに対しHDDが1,000倍程度に留まる：HDDのシーケンシャルアクセス性能により
- HDDはテープより6桁安い

TBscanに着目：
- DRAMは依然として安価
- HDDとテープが近づく（1桁）

TBscanに基づく1997年との比較

- DRAMとHDDのギャップが拡大している
  - 1997年ではDRAMをHDDの代わりに使うのは13倍安かったが，現在では300倍安い
  - スキャン主体のアプリケーションでもHDDを避けた方がよい
- HDDとテープのギャップが小さくなってきている
  - 1997年ではHDDはテープより70倍安かったが，現在は10倍

何が言えるか

- これまでHDDをcapacity tier，テープをarchival tierとして用いてきたが，HDDとテープの接近により，cold storage tierとまとめるのがよい
- 新たなコールドストレージの技術（高速）も出現
  - MAID (massive array of idle disks)
    - 待機時にHDDの電源を停止
- 遅延を気にしないバッチ処理はコールドストレージ上で

https://www.fujitsu.com/jp/Images/fig01_tcm102-1971734.gif

目次

- ジム・グレイについて
- 5分間ルールとは
- ルールの導出
- ルールの適用
- Performance Tierに対する分析
- Capacity Tierに対する分析
- 結論

結論①

- **HDDはもはやテープである**
- 5分間ルールは1987年のDRAM-HDDについて成り立ったが，現在は4時間ルールとなっている
- HDDは性能が重要な場面だけでなく，非シーケンシャルなデータアクセスパターンを持つすべての応用で不適合になりつつある

結論②

- **NVMはもはやDRAMである**
- DRAMとSSDのギャップが縮まる傾向にある
- 現在，5分間ルールはDRAM-SSDの間で成り立つ
- 最新のNVM（3D-XPoint）技術を用いると，1分未満となる
- 今後，DRAMに基づくデータベースエンジンが，NVMに基づく永続メモリエンジンに移行する

結論③

- **コールドストレージはもはや「ホット」である**
- シーケンシャルなワークロードに対し，HDDとテープのギャップが急速に縮まっている
- 最新のコールドストレージ機器はさらに優れる
- HDDは性能重視でないバッチ的な分析にも不適合となっていく：コールドストレージ上で直接実行する方がよい