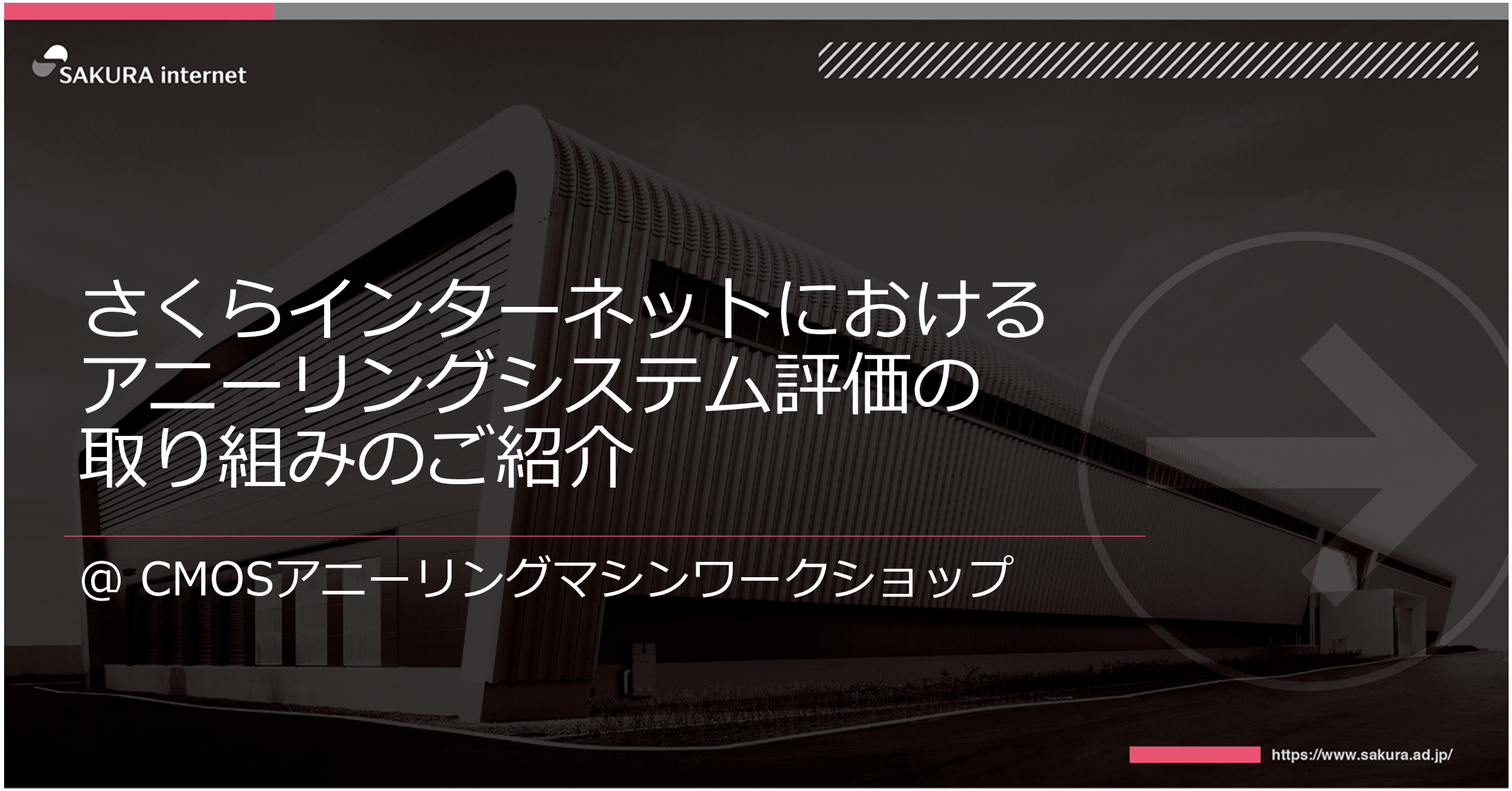
SAKURA internet

# さくらインターネットにおける アニーリングシステム評価の 取り組みのご紹介

## @ CMOSアニーリングマシンワークショップ



https://www.sakura.ad.jp/


DAY
2019/08/30

COMPANY
さくらインターネット株式会社

DEPARTMENT
さくらインターネット研究所

NAME
上級研究員　菊地 俊介

- さくらインターネット、同研究所について
- さくら（研究所）におけるCMOSアニーリングマシン評価・検討の位置づけ
- 評価結果
- アニーリングマシン活用事例のご紹介
- 今後の展開予定

- 本資料は、さくらインターネット研究所ブログで公開予定です。
  - https://research.sakura.ad.jp/

# 菊地 俊介　（東京都出身、品川区在住）

@kikuzokikuzo

https://note.mu/kikuzokikuzo

https://www.facebook.com/kikuzokikuzo

**所属**　さくらインターネット研究所

**学歴**　早稲田大学大学院 理工学研究科
　電子・情報通信学専攻 修士課程修了
早稲田大学大学院 国際情報通信研究科 博士課程単位取得退学

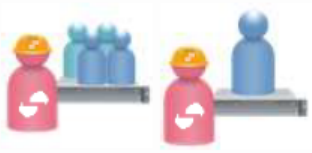
**職歴**　富士通（株）富士通研究所に就職
　ネットの研究やったり、SEやったり、
　NICTに出向したり、トイレIoT作ったり
さくらインターネットに転職
　データ流通実証実験、OpenFogコンソーシアム、
　AR/VR、量子（アニーリング）コンピュータ

**専門**　エッジ・Fogコンピューティング
（分散系システムのあたり）

**趣味**　新技術調査、家庭内IoT、車、鉄道

# 新しい社会のインフラを支えながら、最先端のサービスを構築してゆく

## レンタルサーバ



さくらのレンタルサーバ
さくらのマネージドサーバ

1台のサーバを複数の契約者でサーバを共有または占有することができ、管理はさくらインターネットに任せて使うサービス

1台を共有

1台を占有

## VPS


さくらのVPS

仮想化技術を用い、1台の物理サーバ上に複数の仮想サーバを構築し、仮想専用サーバとして分けた領域の占有サービス

## クラウド

さくらのクラウド
SAKURA CLOUD

高性能サーバと拡張性の高いネットワークを圧倒的なコストパフォーマンスで利用できるIaaS型パブリック・クラウド・サービス

## 専用サーバ

さくらの専用サーバ

高性能で拡張性と信頼性の高いサーバをまるごと独占して利用することができ、自由にカスタマイズして利用可能なサービス

1台～複数台

## データセンター

ハウジング
リモートハウジング

データセンター内にお客様専用のハウジングスペースを確保し、ネットワーク機器やサーバなどの機材を自由に置けるサービス

## 新サービス

sakura.io

通信環境とデータの保存や処理システムを一体型で提供するIoTプラットフォーム・サービス

https://sakura.io/

Arukas
HOSTING DOCKER CONTAINERS

Dockerコンテナをマネージドされた環境へ、手軽・シンプルにプロビジョニング可能なサービス

https://arukas.io/

高火力
KOUKARYOKU

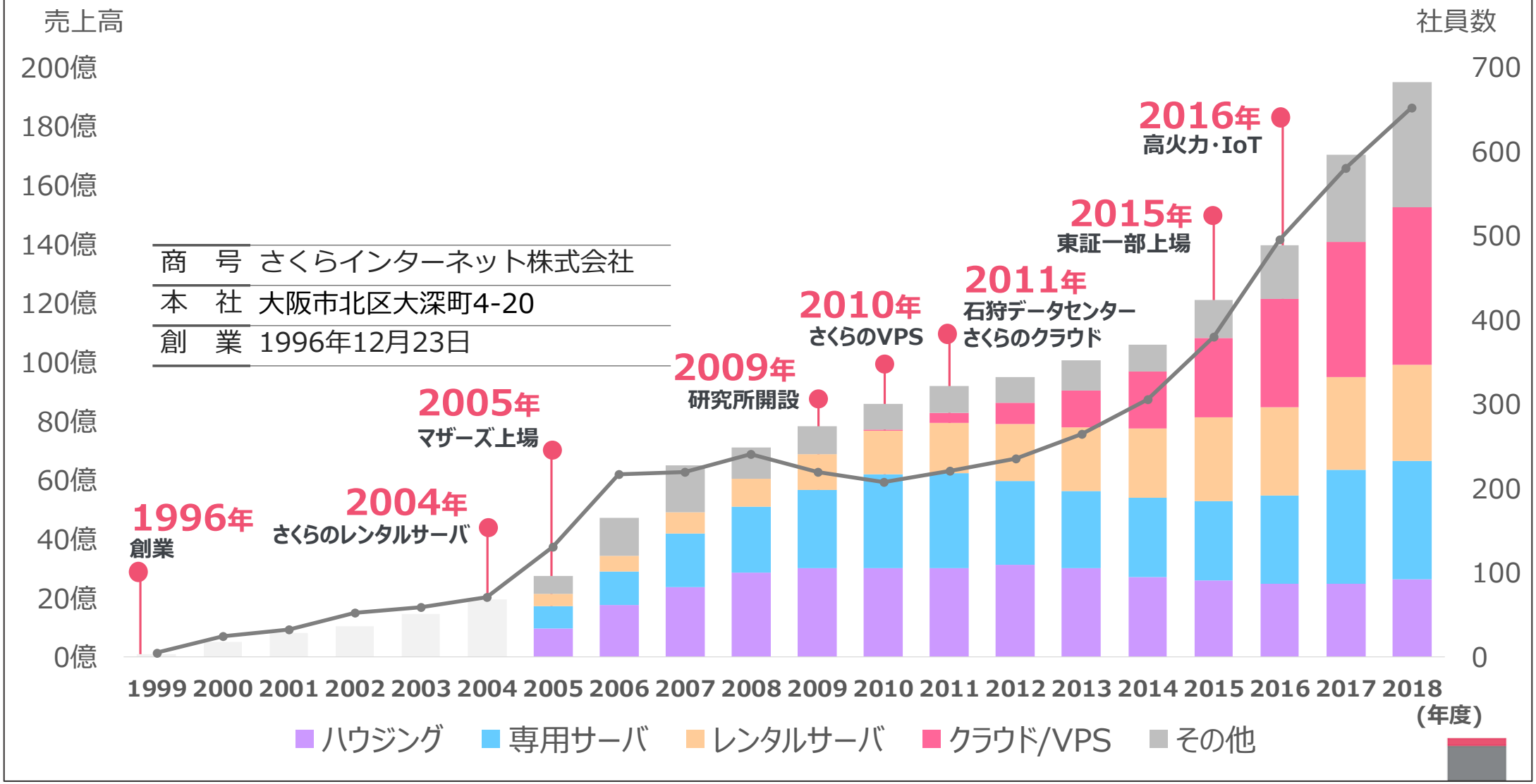
高火力コンピューティング

https://www.sakura.ad.jp/koukaryoku/

## サービスの主な利用用途

- ウェブサイト運営、ブログ、インターネット・メール
- ネットビジネス、電子商取引、動画・音楽配信、開発環境
- エンタープライズ
- 会員制サイト、キャンペーン・サイト
- SNS、ウェブ・アプリケーション、SaaS、ASP

## 沿革




| | |
|---|---|
| 商　号 | さくらインターネット株式会社 |
| 本　社 | 大阪市北区大深町4-20 |
| 創　業 | 1996年12月23日 |

SAKURA internet

# 高火力
KOUKARYOKU

## AI・ディープラーニングに最適な高火力 GPU サーバー

高火力コンピューティング ® とは、コンピューティングリソースの運用や設備投資のリスクを省き、AIやディープラーニングの性能を追求して競争力を高めたい方のニーズにお応えする、さくらインターネットの計算リソースサービスです。

ビジネスニーズに最適
GPUサーバーは「所有」から「利用」へ

導入しやすいお手頃モデルから
最新モデルまで、幅広いラインアップ

月額課金と時間課金
最適な料金プランを選択可能

おすすめ用途：機械学習 / ディープラーニング / 画像処理 / スパコン / レンダリング

## Tesla V100（32GB）モデルを提供開始

従来モデルの10倍以上相当となる100TFLOPS を超える性能を持つV100モデル。GPU搭載メモリ16GBの価格改定に加え、32GBを提供開始いたしました。

| | |
|---|---|
| 単精度浮動小数点数演算性能 | 約14TFLOPS |
| 倍精度浮動小数点数演算性能 | 約7TFLOPS |
| Tensor コア | 約112TFLOPS |
| GPU搭載メモリ | 32GB HBM2メモリ搭載 |
| オプションサービス | CPU・メモリ・SSD・GPU追加<br>回線アップグレード |
| おすすめの用途 | 深層学習（ディープラーニング）<br>科学技術計算 |
| 月額利用料（税込） | 124,200円 |
| 初期費用（税込） | 1,010,880円 |

**ご検証いただくための無償トライアルユーザーも募集しております。**

「さくら　高火力」で検索　website:　https://www.sakura.ad.jp/koukaryoku/

さくらの専用サーバ
高火力シリーズ

「高火力コンピューティング」では主に機械学習用途にご利用いただけるGPUサーバーを４モデルご用意しております。
NVIDIAのデータセンター向けGPU、Tesla P40またはP100に加え、ディープラーニングのために設計された新しい演算器アレイ「Tensorコア」が搭載されたV100モデルも選択いただけます。

| | Tesla P40 モデル | Tesla P100 モデル | Tesla V100(16GB)モデル | Tesla V100(32GB)モデル |
|---|---|---|---|---|
| 時間課金<br>利用料金 | 1時間あたり349円<br>（税込：376円） | 1時間あたり357円<br>（税込：385円） | – | |
| 月額課金<br>利用料金 ※1 | 月額97,000円<br>（税込：104,760円）<br>初期費用875,000円<br>（税込：945,000円） | 月額99,000円<br>（税込：106,920円）<br>初期費用895,000円<br>（税込：966,600円） | 月額99,000円<br>（税込：106,920円）<br>初期費用924,000円<br>（税込：997,920 円） | 月額115,000円<br>（税込：124,200円）<br>初期費用936,000円<br>（税込：1,010,880 円） |
| GPUカード | NVIDIA Tesla P40 x1 | NVIDIA Tesla P100 x1 | NVIDIA Tesla V100 for PCI-Express (16GB) x1 | NVIDIA Tesla V100 for PCI-Express (32GB) x1 |
| GPU搭載メモリ | 24GB | 16GB HBM2メモリ搭載 | 16GB HBM2メモリ搭載 | 32GB HBM2メモリ搭載 |
| 単精度浮動小数点数演算性能 | 約12.0 TFLOPS | 約9.3 TFLOPS | 約14 TFLOPS | |
| 倍精度浮動小数点数演算性能 | – | 約4.7 TFLOPS | 約7 TFLOPS | |
| Tensor コア | – | – | 約112 TFLOPS | |

| 基本スペック | 標準 | オプション ※2 |
|---|---|---|
| CPU | Xeon E5-2623 v3 4コア × 2（8C/16T 3.0GHz 最大3.5GHz） | |
| メモリ | 128GB | 最大1TBまで増設可能 |
| ストレージ | SSD 480GB 2台/1組（RAID1） | 最大8台まで増設可能（標準2台＋6台追加可能） |
| グローバル回線 | 100Mbps 冗長構成、ベストエフォート | 帯域追加、優先制御 |
| ローカル回線 | 10Gbps 冗長構成、ベストエフォート | InfiniBand インターコネクト（個別御見積） |
| OS | 標準OSは「Ubuntu 16.04 (64bit)」、「Ubuntu 14.04 (64bit)」「CentOS 7」をユーザー自身でインストール可能 | |

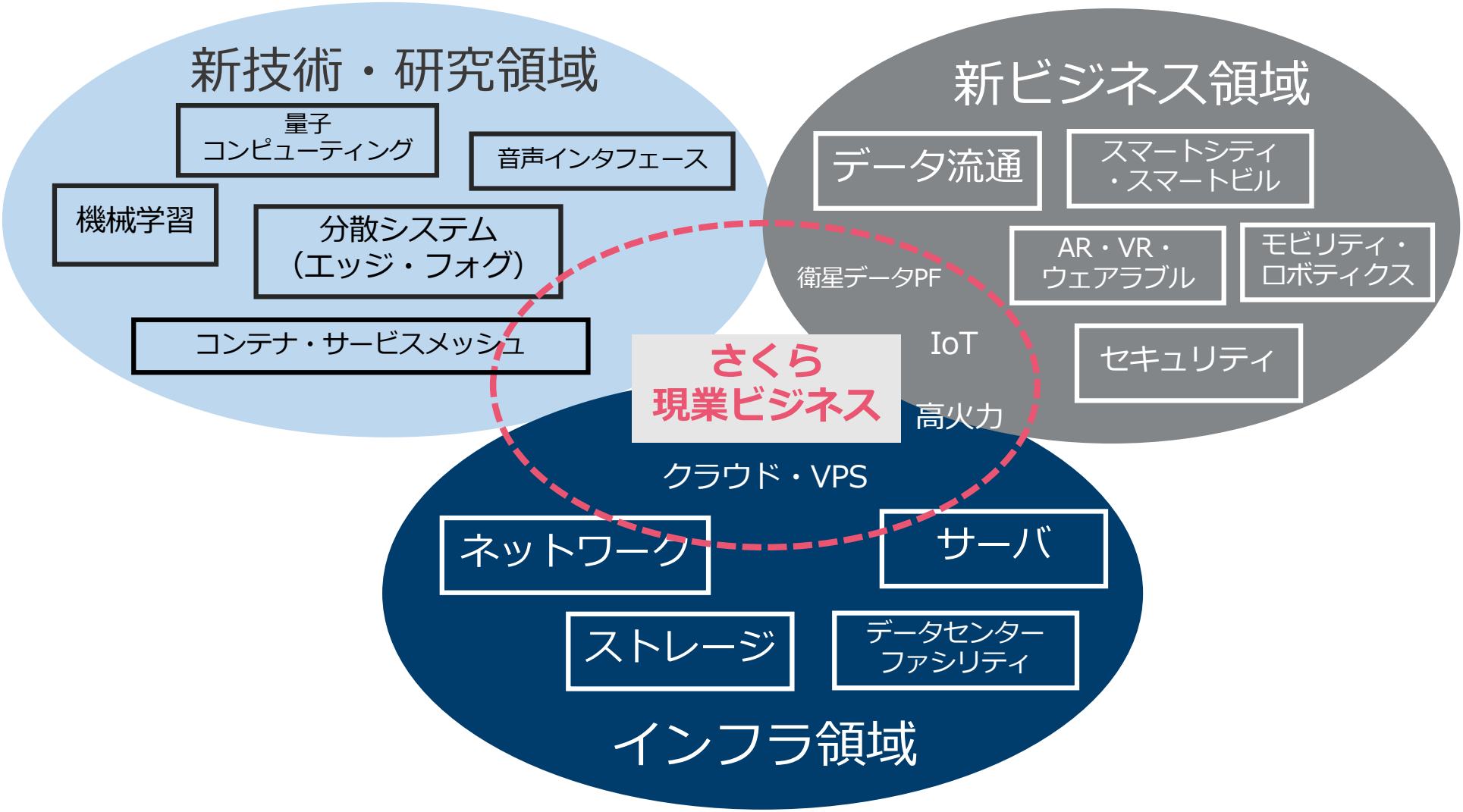
※1 最低利用期間について、初期費用一括払いの場合は3ヶ月、分割払いの場合は12ヶ月となります。また年間一括払いで1ヶ月分お得になります。
※2 オプションは月額課金のみご利用いただけます。



さくらインターネット株式会社
［大阪本社］〒530-0011 大阪府大阪市北区大深町4-20 グランフロント大阪タワーA 35F
［東京支社］〒160-0023 東京都新宿区西新宿7-20-1 住友不動産西新宿ビル3号館
［福岡オフィス］〒810-0042 福岡県福岡市中央区赤坂1-12-15 読売福岡ビル 7F
0120-380397
受付時間：平日10:00～18:00（土日・祝祭日を除く）
E-MAIL support@sakura.ad.jp

高火力コンピューティングの詳細
https://www.sakura.ad.jp/koukaryoku/

# さくらの現業ビジネスの先にあるものを見ていく。



SAKURA internet

Press Release

報道関係各位

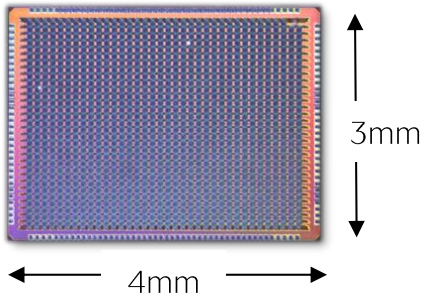
2018 年 10 月 18 日
さくらインターネット株式会社

# さくらインターネット、組合せ最適化問題の高速処理を実現する 日立製作所の新概念コンピューター「CMOS アニーリングマシン」の評価を開始 ～評価の一環で「CMOS アニーリングマシン」の未来を議論するセミナーを開催～

インターネットインフラサービスを提供するさくらインターネット株式会社（本社：大阪府大阪市、代表取締役社長：田中 邦裕）の組織内研究所であるさくらインターネット研究所は、CPU や GPU に次ぐ新たな計算機である日立製作所の新概念コンピューター「CMOS アニーリングマシン」※1の評価を開始します。また、評価の一環として、技術者間で「CMOS アニーリングマシン」の未来について議論し、新たなニーズやビジネス化の可能性を模索するセミナーを開催します。

■CMOS アニーリングマシン

・第 1 世代アニーリングマシン




・大規模 FPGA 構築イメージ

# CMOSアニーリング計算機セミナー

*CMOS Annealing Computer Seminar*

## 第一回CMOSアニーリング計算機セミナーサイトへようこそ

インターネットインフラサービスを提供するさくらインターネット株式会社（本社：大阪府大阪市、代表取締役社長：田中 邦裕）の組織内研究所であるさくらインターネット研究所は、CPUやGPUに次ぐ新たな計算機である日立製作所の新概念コンピューター「CMOSアニーリングマシン」の評価を開始します。また、評価の一環として、技術者間で「CMOSアニーリングマシン」の未来について議論し、新たなニーズやビジネス化の可能性を模索するセミナーを開催します。

交通渋滞の解消や物流コストの最小化、電力送電網による安定したエネルギー供給など、複雑化する社会システムの課題解決には、全体最適となる組合せを見いだすことが重要です。しかし、社会システムが複雑化すると、システムを記述するパラメータとその組合せが爆発的に増大し、最適なパラメータを決定することが困難になります。この度、さくらインターネット研究所では、これら組合せ最適化問題※2を解くにあたりCPUやGPUを搭載する一般的なコンピューターとは異なる新たなコンピューター「CMOSアニーリングマシン」に着目し、組合せ最適化問題の高速処理やデータセンターでの利用に向けた評価を開始します。また、評価の一環で技術者を対象に「CMOSアニーリングマシン」の新たなニーズやビジネス化の可能性の議論と意見交換の場としてセミナーを開催します。

セミナー開催概要

- 日時　２０１８年１１月２２日（木曜）　１０：００-１８：００
- 場所　さくらインターネット株式会社　西新宿セミナールーム
  - 東京都新宿区西新宿４丁目３３-４　住友不動産西新宿ビル４号館５F
- プログラム概要
  - 午前　CMOSアニーリングマシンの概要説明
  - １１：３０-１３：００　昼食休息（各自）
  - 午後　実機評価およびディスカッション
- 参加条件
  - CMOSアニーリングマシンに興味があり、積極的な利用検討が行える方。

## 次のさくらのビジネスを支える（かもしれない）要素技術へのキャッチアップ

これまでも、新しい技術に支えられた新しいサービスが、会社を成長させてきた：「クラウド」「高火力」

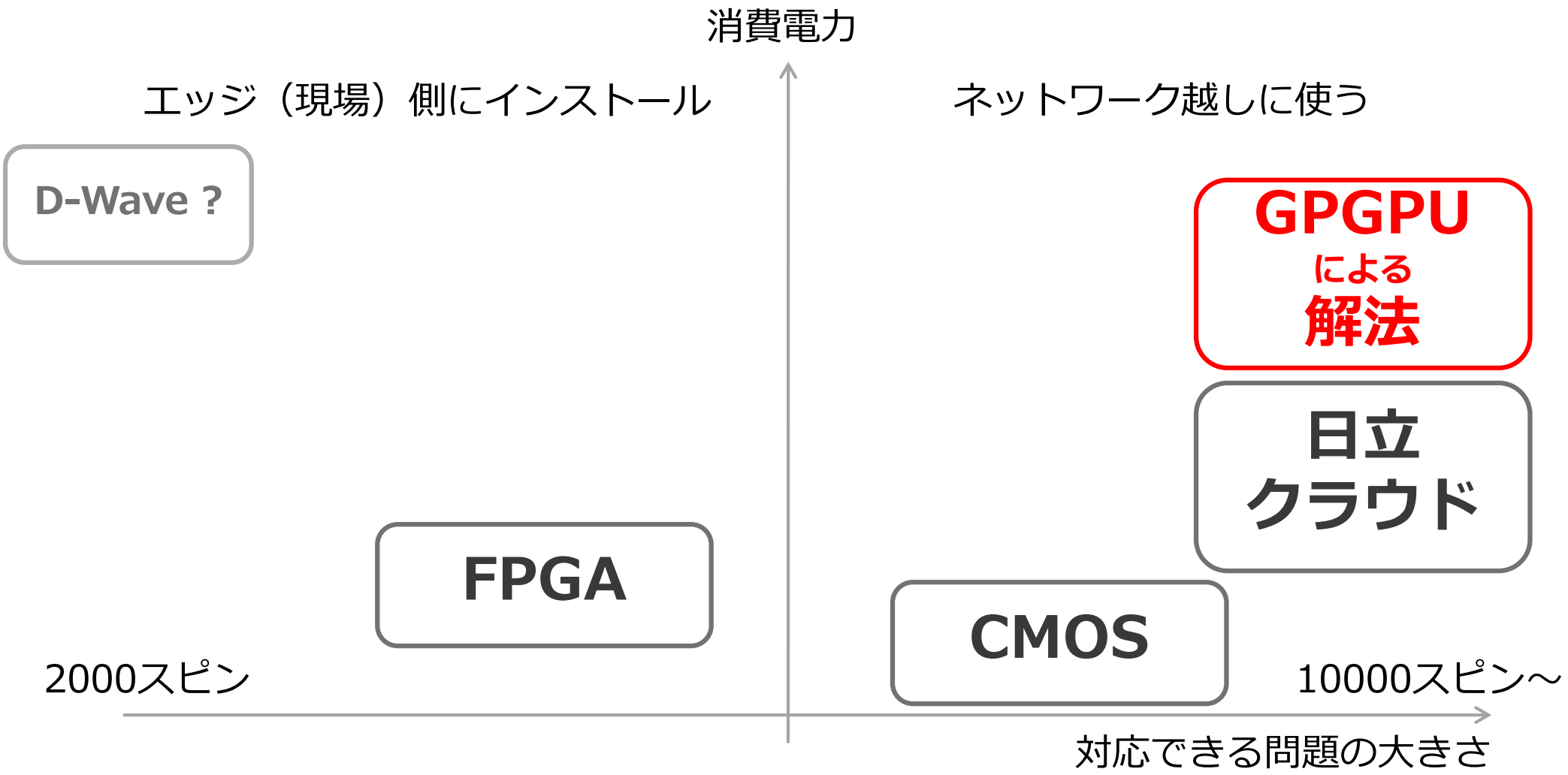
## 新しい需要の創出への支援

NVIDIAが様々なGPU関連製品を出し、使い方セミナーを開催するのと同様。その立場にある日立さんの支援。

## 自社内サービスへの適用の可能性

（後ほど紹介）

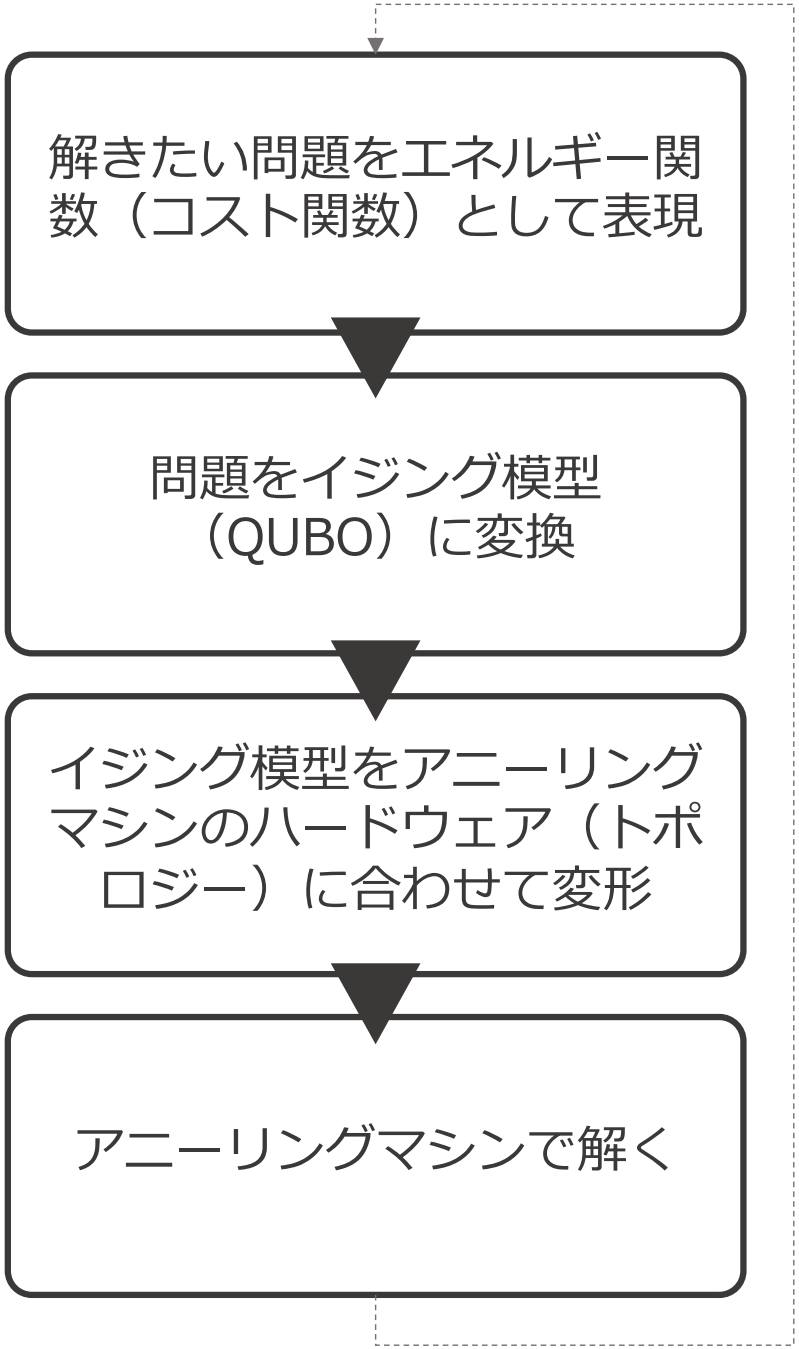
# ソリューションとしての位置づけ



# ビジネスソリューションの構成（想定、可能性）

- フルサポート・フルマネージドサービスとして：
  - 日立さんによる対応

- 自社開発向けソリューションとして：
  - さくらで環境準備のお手伝い？




pyQUBOを利用する

今回の評価では、ランダムグラフを利用

ソフトウェア定義でフルメッシュ結合のため、変形不要

**さくらの高火力環境で評価**

# 問題を定式化する

（アニーリングマシンを使う上で一番難しいネックとなるポイントがここ）

$$E(\{s_i\}) = \sum_{i \in V} h_i s_i + \sum_{(i,j) \in E} J_{ij} s_i s_j$$

数式は日立サイト（https://annealing-cloud.com/tutorial/1.html）より借用

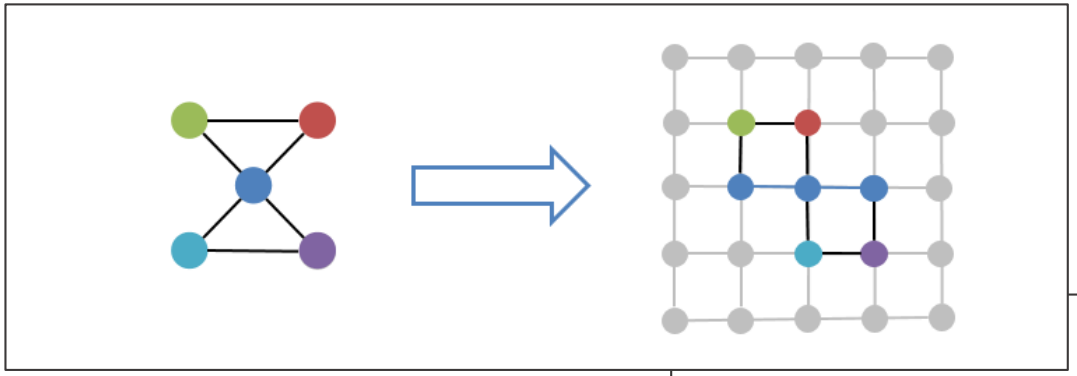
（ここはツールなどのソリューションがあまりなく、自力でやらなければならないところ）

# 参考サイト

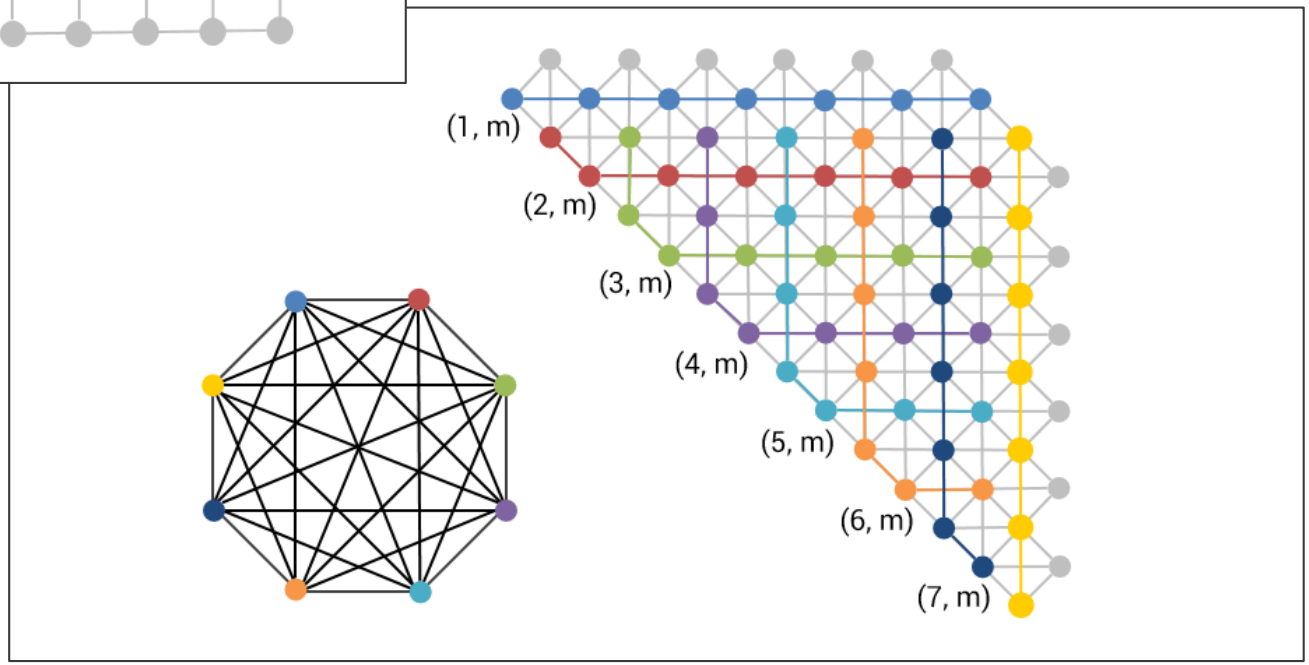
- QUANTUM COMPUTING SOLUTIONS：フィックスターズ社の解説サイト
  - https://quantum.fixstars.com/introduction_to_quantum_computer/quantum_annealing/ising_model/
- T-Wave：東北大学量子アニーリング研究センターが運営するナレッジベース
  - https://qard.is.tohoku.ac.jp/T-Wave/?p=43

# イジング模型を、アニーリングマシンのハードウェア構成に合わせて変形する必要がある

（今回のソリューションでは、フルメッシュ結合となっているため、対応不要。）（すばらしい！）



キンググラフの場合




図はいずれも
https://quantum.fixstars.com/introduction_to_quantum_computer/quantum_annealing/programming/graph_mapping/ より借用

さくらの「高火力コンピューティング」（GPGPU搭載ホスト時間貸し環境）でMAソフトウェアを使用した場合の、使い勝手を見出すことを目的に評価

## 評価した内容

- 評価1: 計算量と計算時間の特性
- 評価2: 精度の特性

## 費用目安

今回の評価の対象外

- 性能の上限
- 既存手法（SAや日立製作所の既存ソリューション）との比較

- さくらインターネット高火力コンピューティング環境を利用

今回評価対象

**Tesla P40モデル**

NVIDIA Tesla P40 x1
- 24GB メモリ
- 単精度性能：約12.0TFLOPS
- 倍精度性能：- (実装なし)

**Tesla P100モデル**

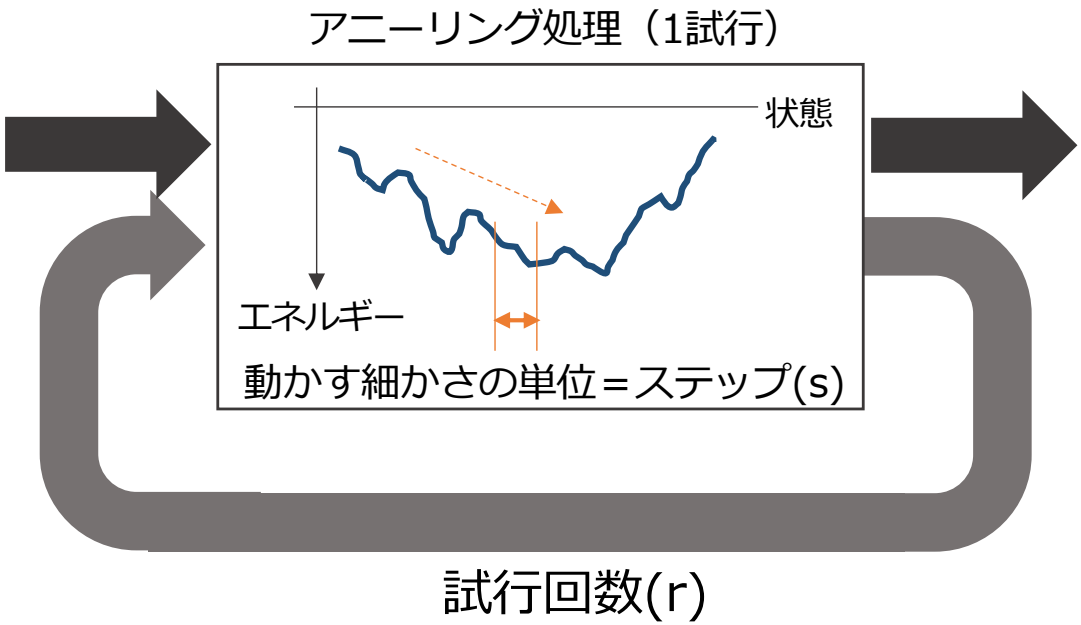
NVIDIA Tesla P100 x1
- 16GB メモリ
- 単精度性能：約9.3 TFLOPS
- 倍精度性能：約4.7 TFLOPS

- 入力イジングモデルとして、ランダムグラフを利用
  - ランダムグラフ生成には”rudy”を使用
    - 汎用グラフジェネレータとして広く用いられている
    - https://web.stanford.edu/~yyye/yyye/Gset/

# 評価1-1:一つのイジングモデルに対して、アニーリング処理のステップ数、試行回数を変化させて計算時間を観測

## （日立製）アニーリング処理の構成

イジングモデル、ステップ数(S)、試行回数(R)をパラメータとして与え、計算開始




試行回数分の結果をまとめて（ヒストグラム化、平均値等を計算等）、出力とする

# 評価1-1:一つのイジングモデルに対して、アニーリング処理のステップ数、試行回数を変化させて計算時間を観測

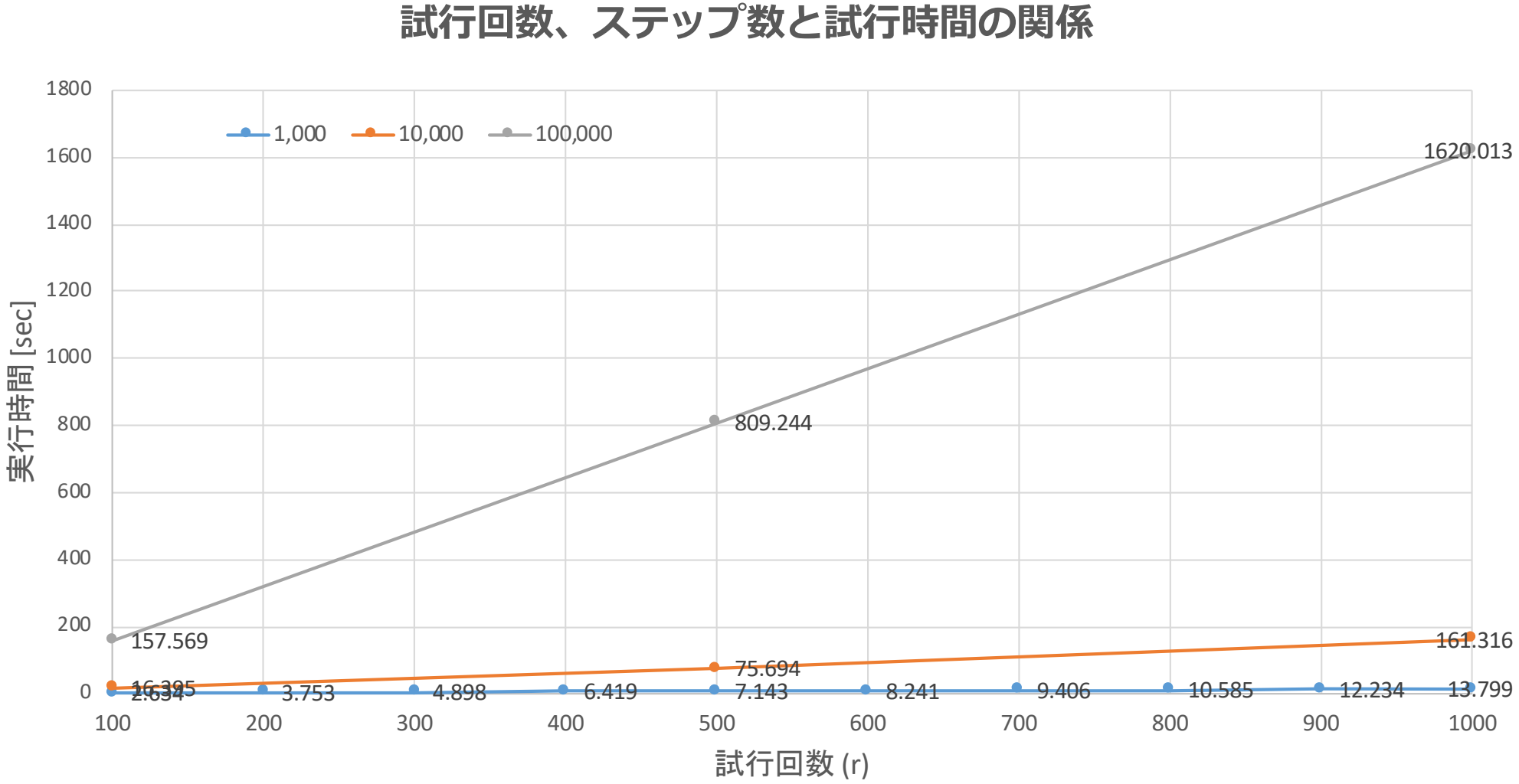
d=200

| ステップ数 (s) | 試行回数 (r) 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1,000 | 9.04 | 8.88 | 8.69 | 9.34 | 8.66 | 8.61 | 8.56 | 8.6 | 8.98 | 9.35 |
| | 904 | 1,776 | 2,607 | 3,736 | 4,330 | 5,166 | 5,992 | 6,880 | 8,082 | 9,350 |
| | **2.634** | **3.753** | **4.898** | **6.419** | **7.143** | **8.241** | **9.406** | **10.585** | **12.234** | **13.799** |
| 10,000 | 122.11 | - | - | - | 121.02 | - | - | - | - | 130.96 |
| | 12,211 | - | - | - | 60,510 | - | - | - | - | 130,960 |
| | **16.395** | - | - | - | **75.694** | - | - | - | - | **161.316** |
| 100,000 | 1286.11 | - | - | - | 1335.53 | - | - | - | - | 1337.82 |
| | 128,611 | - | - | - | 667,765 | - | - | - | - | 1,337,820 |
| | **157.569** | - | - | - | **809.244** | - | - | - | - | **1620.01** |

凡例：

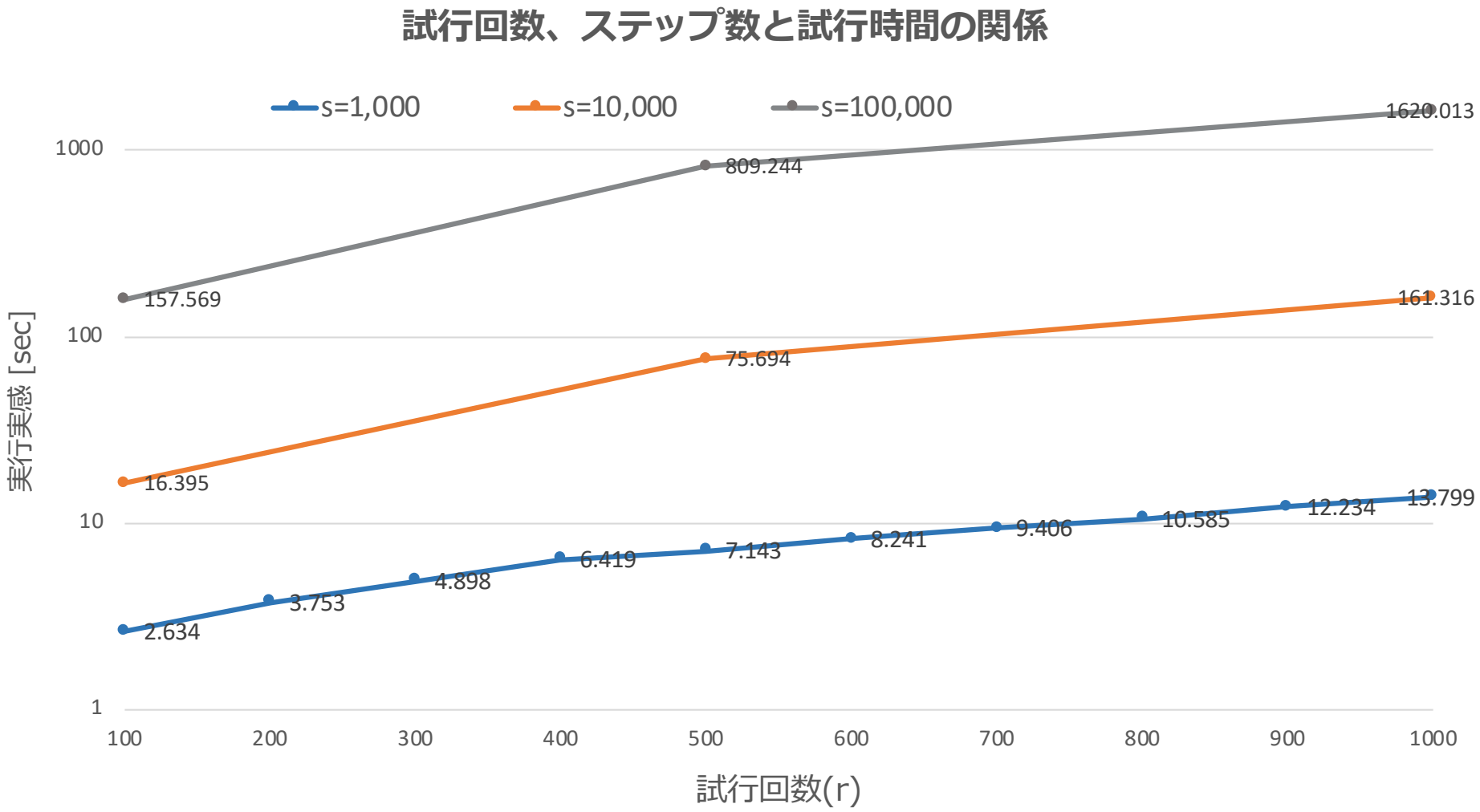
| |
|---|
| 1試行あたり [ms] |
| 全試行合計 [ms] |
| **コマンド実行時間 [s]** |

# 評価1-1:一つのイジングモデルに対して、アニーリング処理のステップ数、試行回数を変化させて計算時間を観測




**ステップ数、試行回数に対して計算時間はリニアに変化する。**

# 評価1-1:一つのイジングモデルに対して、アニーリング処理のステップ数、試行回数を変化させて計算時間を観測




**ステップ数あたりの効率は、試行回数が増えると悪化する。**

SAKURA internet

# 評価1-2:入力イジングモデルのサイズ（スピン数）をパラメータに、試行回数を変化させて計算時間を観測

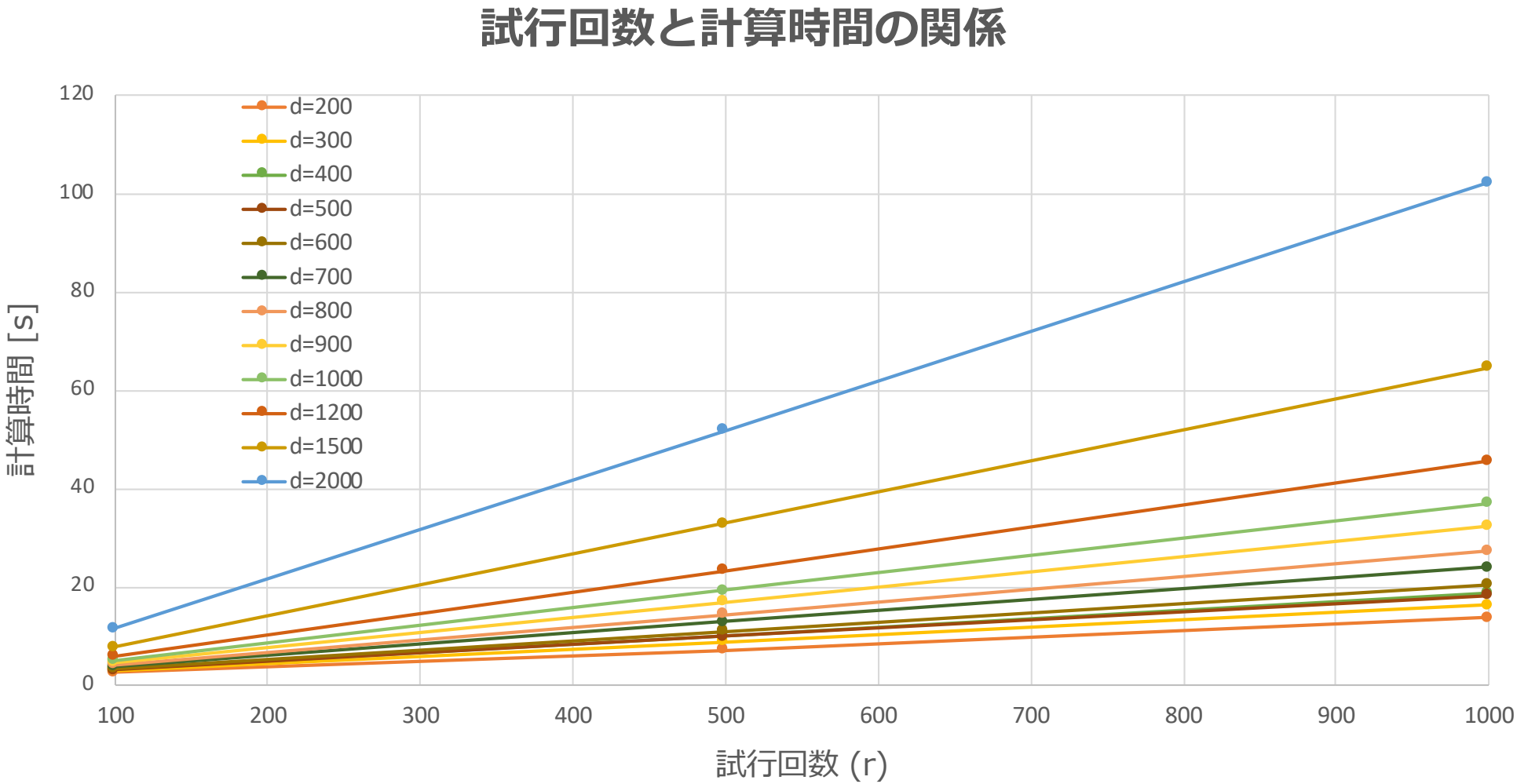
s=1000

| 試行回数(r) | スピン数(d) 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 | 1200 | 1500 | 2000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 9.04 | 11.5 | 13.87 | 12.64 | 13.83 | 16.65 | 18.91 | 23.07 | 27.03 | 33.39 | 49.14 | 79.86 |
| | 904 | 1,150 | 1,387 | 1,264 | 1,383 | 1,665 | 1,891 | 2,307 | 2,703 | 3,339 | 4,914 | 7,986 |
| | **2.634** | **2.914** | **3.239** | **3.158** | **3.347** | **3.767** | **4.088** | **4.582** | **4.994** | **5.877** | **7.82** | **11.615** |
| 500 | 8.66 | 11.27 | 13.74 | 12.39 | 13.82 | 16.57 | 18.75 | 22.98 | 26.98 | 33.25 | 48.93 | 79.84 |
| | 4,330 | 5,635 | 6,870 | 6,195 | 6,910 | 8,285 | 9,375 | 11,490 | 13,490 | 16,625 | 24,465 | 39,920 |
| | **7.143** | **8.891** | **10.111** | **9.91** | **11.028** | **12.816** | **14.422** | **16.855** | **19.201** | **23.415** | **32.902** | **51.856** |
| 1000 | 9.35 | 11.32 | 13.63 | 12.26 | 13.6 | 16.48 | 18.85 | 23.04 | 26.88 | 33.47 | 49.08 | 79.95 |
| | 9,350 | 11,320 | 13,630 | 12,260 | 13,600 | 16,480 | 18,850 | 23,040 | 26,880 | 33,470 | 49,080 | 79,950 |
| | **13.8** | **16.354** | **18.719** | **18.174** | **20.385** | **24.061** | **27.303** | **32.347** | **36.917** | **45.598** | **64.522** | **102.24** |

凡例：

| |
|---|
| 1試行あたり [ms] |
| 全試行合計 [ms] |
| **コマンド実行時間 [s]** |

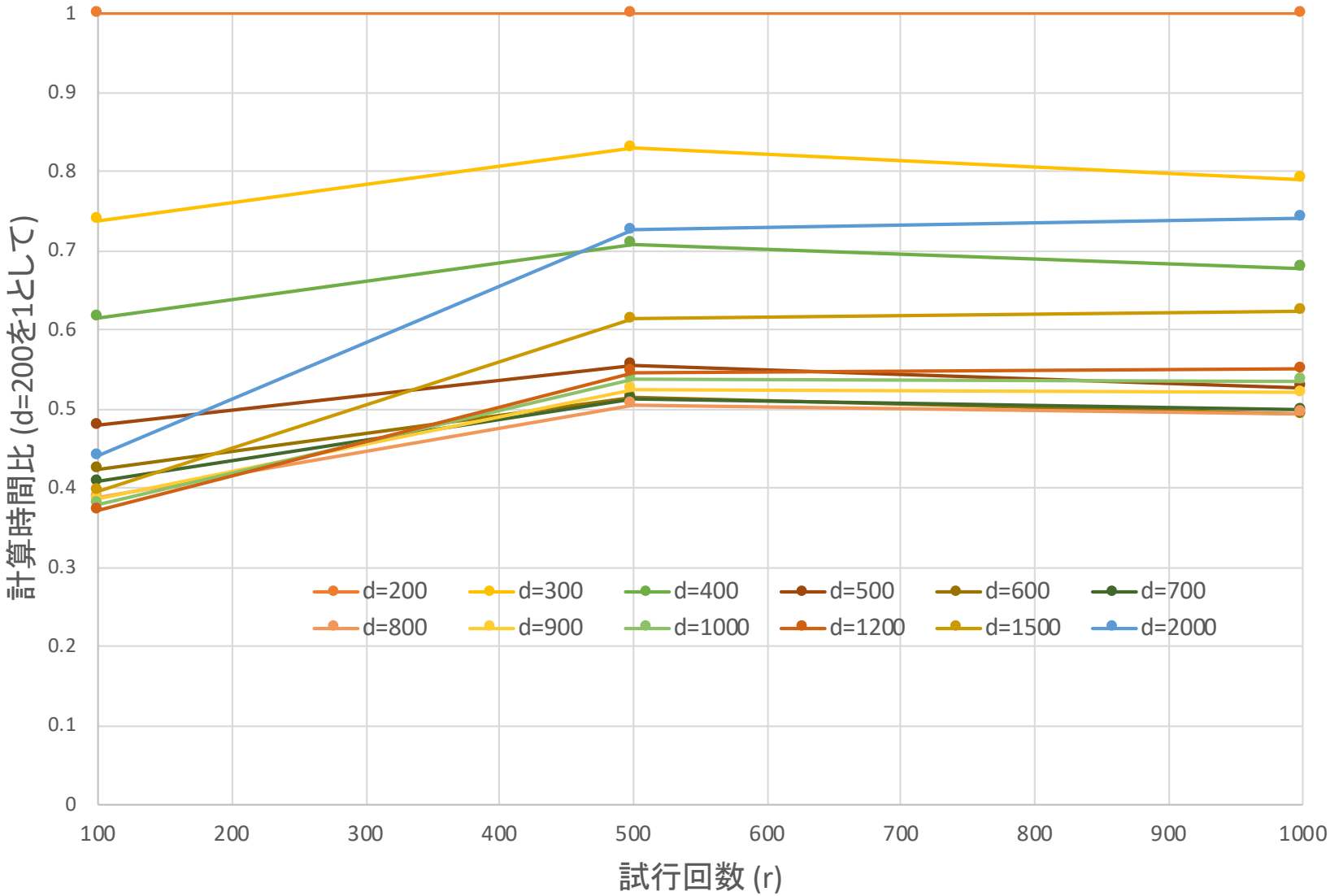
# 評価1-2:入力イジングモデルのサイズ（スピン数）をパラメータに、試行回数を変化させて計算時間を観測




**試行回数に対して計算時間はリニアに変化する。**
**それはスピン数を変えても変わらない。**

評価1-2:入力イジングモデルのサイズ（スピン数）をパラメータに、試行回数を変化させて計算時間を観測

## スピン数と計算時間の関係（比率）




**スピン数が大きくなると計算効率は上がる**

# 評価2:d=200のイジングモデルにおいて、試行回数とステップ数を変化させて最小エネルギー到達までの変化を確認

d=200　平均エネルギー値と最小エネルギー値

| ステップ数 (s) | 試行回数 (r) 100 | 500 | 1000 |
|---|---|---|---|
| 1,000 | -1975.1 | -1977.13 | -1979.35 |
| | -2042 | -2052 | -2058 |
| 10,000 | -2012.4 | -2014.35 | -2013.67 |
| | -2058 | -2058 | -2058 |
| 100,000 | -2027.92 | -2031.26 | -2031.54 |
| | -2058 | -2058 | -2058 |

凡例：平均エネルギー値 / 最省エネルギー値

d=200　【参考】計算時間

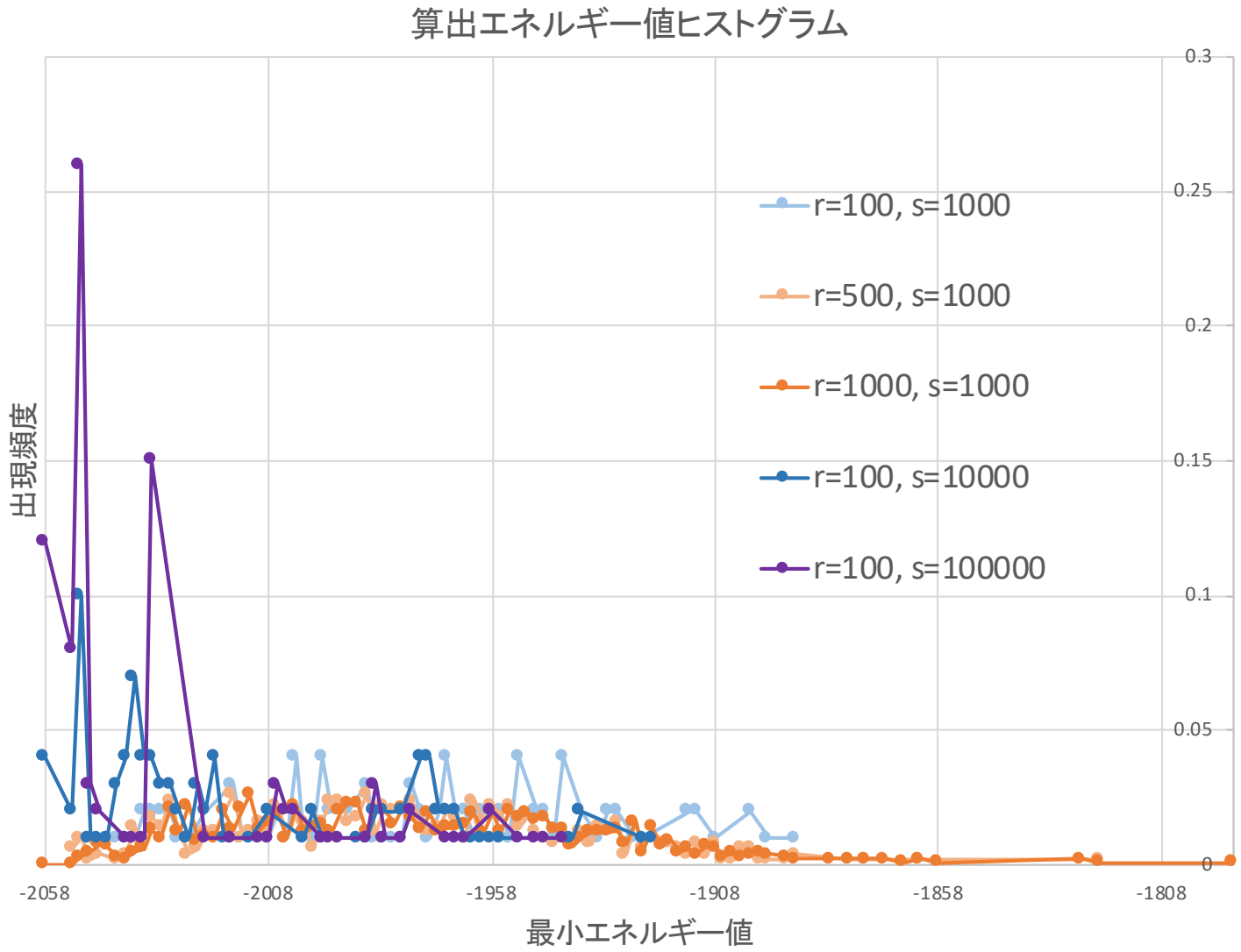
| ステップ数 (s) | 試行回数 (r) 100 | 500 | 1000 |
|---|---|---|---|
| 1,000 | 9.04 | 8.66 | 9.35 |
| | 904 | 4,330 | 9,350 |
| | **2.634** | **7.143** | **13.799** |
| 10,000 | 122.11 | 121.02 | 130.96 |
| | 12,211 | 60,510 | 130,960 |
| | **16.395** | **75.694** | **161.316** |
| 100,000 | 1286.11 | 1335.53 | 1337.82 |
| | 128,611 | 667,765 | 1,337,820 |
| | **157.569** | **809.244** | **1620.013** |

（ステップ数・試行回数と計算時間の関係表より抜粋）

凡例：1試行あたり [ms] / 全試行合計 [ms] / **コマンド実行時間 [s]**

**試行回数よりもステップ数を増加させたほうが、求解には有利**

# 評価2:d=200のイジングモデルにおいて、試行回数とステップ数を変化させて最小エネルギー到達までの変化を確認




**試行回数よりもステップ数を増加させたほうが求解には有利**

## 特性

- 試行回数とステップ数に関して、計算時間は線形に増加する
- 入力するイジングモデルのサイズに対して、計算時間はほぼ線形に増加する
- むしろ、入力サイズが大きくなるほど効率が上がる

## 計算時間（速度）

- 200ノード、1000ステップ100試行で2.6秒
- 200ノード、1000ステップ1000試行で13.8秒
- 200ノード、10000ステップ100試行で16.4秒
- 1000ノード、1000ステップ1000試行で36.9秒

- **クエリに対してリアルタイムで応答させるようなタイプの処理には（まだギリギリ）向かない。バッチ処理に組み込む分には十分高速。**
- **（SAに対しては1000倍くらい高速な模様。）**

https://journals.aps.org/pre/abstract/10.1103/PhysRevE.100.012111

PHYSICAL REVIEW E

Images

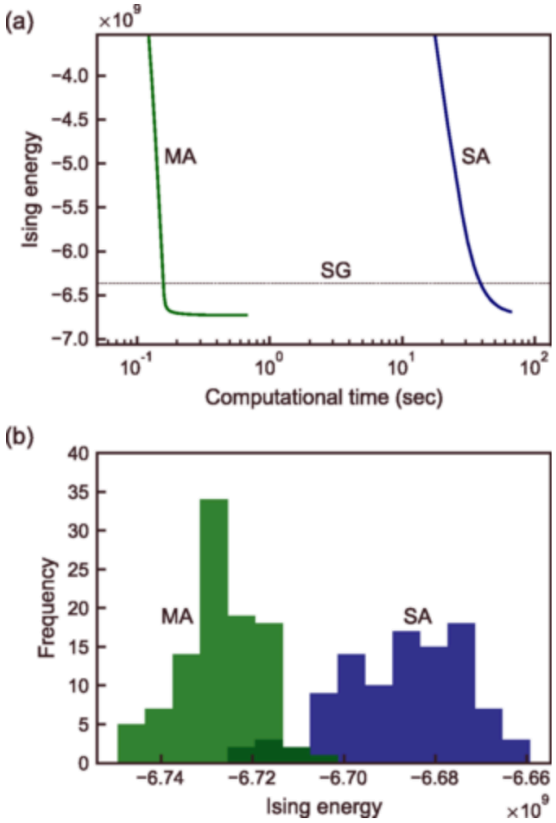



Figure 6

Experimental results obtained with MA and SA solving fully connected Ising model with 100 000 spins and 10 bit widths.
(a) Temporal curve of Ising energy. The dashed line corresponds to the target Ising energy of $-6\,363\,591\,595$ obtained with SG.
(b) Histograms of Ising energy after both methods performed 2000 sweeps.

Reuse & Permissions

Received 3 December 2018 Revised 12 February 2019

Log in with individual APS Journal

# 高火力コンピューティング Tesla P40モデル　の場合

- 時間貸しの場合
  - 環境セットアップ：10分弱
  - アニーリング計算：～15分程度（目安、解かせたい問題サイズに依存）
  - 結果の回収：1分

  - **→349円 (1時間分)**

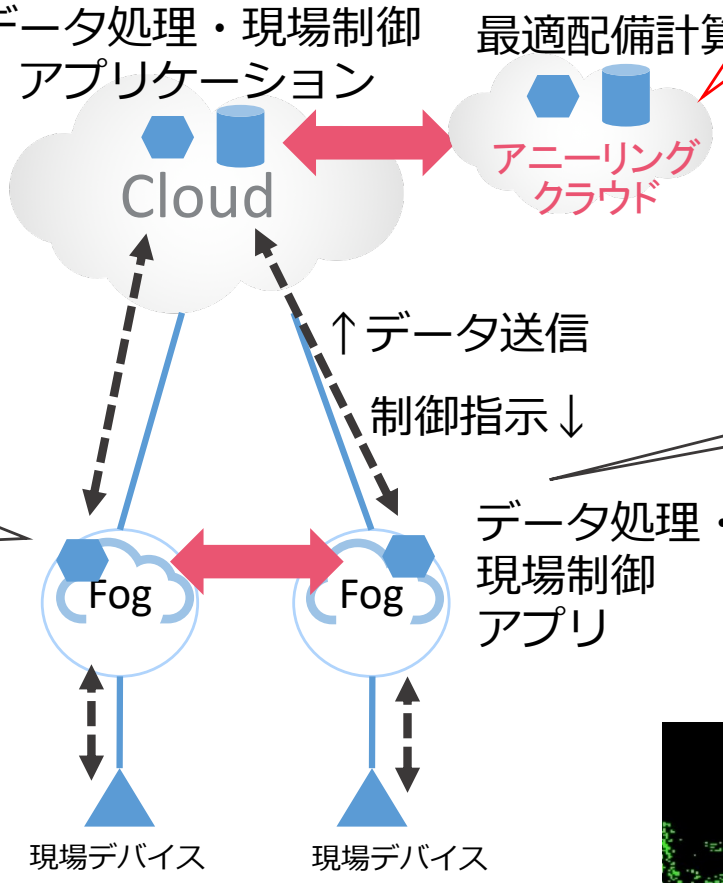
  - 1ヶ月借りっぱなしで**251,280円**

- 月額課金契約で借りる場合（最低3ヶ月以上）
  - 初期費875,000円＋月額97,000円 x3 = **1,166,000円**～

# Fogコンピューティング試作

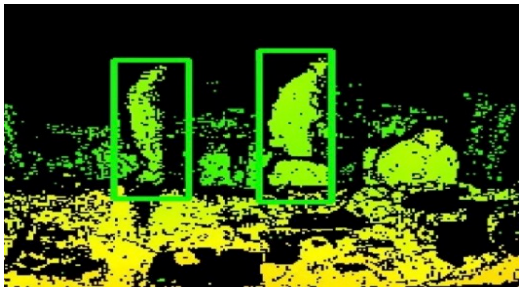
現場とクラウドを連携させて業務改善
（倉庫オペレーション改善を想定）

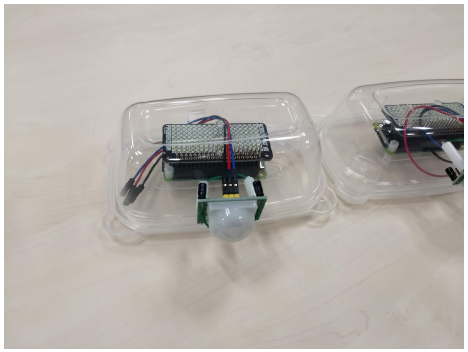
棚への荷物の最適な
格納の仕方を解く




フィールドに現れた
荷物の大きさを検出

ToFカメラ





センサ+LED表示

- さくらのクラウドにおける、VMの配備計画
- エッジノードへのジョブ配備計画
- 画像処理フィルタへの適用

- 検討中…

（案であり本決まりのものではありません）

- アニーリングマシン（GPGPU版）利用トライアルユーザ募集
- アニーリングマシン（GPGPU版）社内適用（VM最適配備への適用）
- エッジコンピューティング・Fogコンピューティング試作（実際に動かして使う、デモシステム）への適用
  - 「**現場で小問題を高速に解く**」で、使っていきながら評価を進める方向性を重視している。

ありがとうございました。