# 京都大学新スーパーコンピュータシステムの構成と導入アプリ検討

深沢 圭一郎

京都大学学術情報メディアセンター

# 京大スパコン

## T2Kの2世代目以降は主に3つのシステムから構成されてきました

システムA

- メニーコア、自作コードを持った研究者がプロダクトランに使うイメージ。

システムB

- Xeon+DDRの汎用計算機、ISVアプリやシングルノードユーザ向け。

システムC

- 大容量メモリ、共有メモリアプリ向け。

システムを一つにすることで全体の理論性能が高くなるとを目指さずに、比較的広くユーザに利用してもらえるようにしています。

# 京大スパコン歴史

## スパコンでは無く汎用コンとしてFACOM-230-60が1969年に稼働したのが始まり

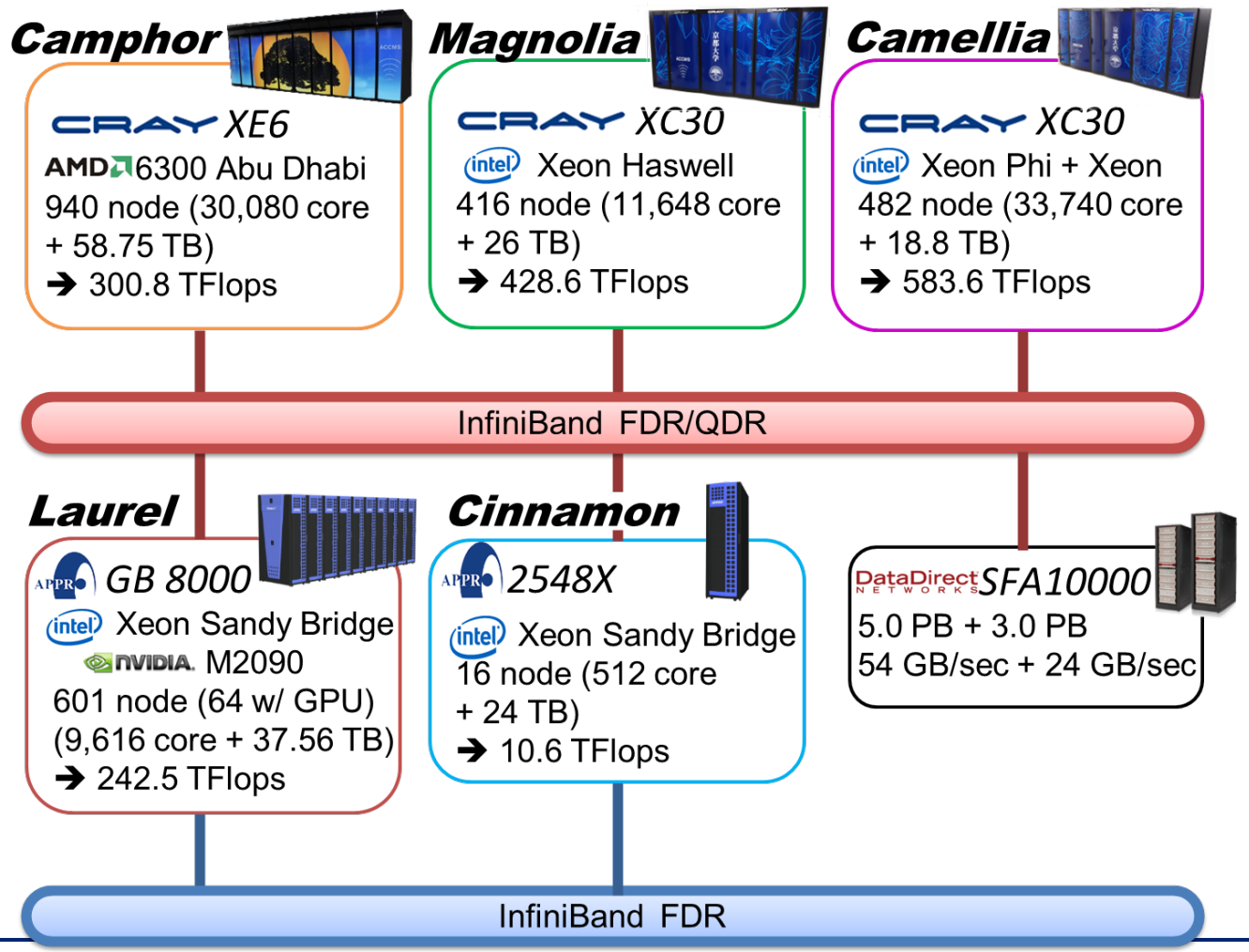
スパコンとしては

- 1985年 FACOM VP100→1986年 VP200+1987年 VP400
- 1991年 Fujitsu VP2600/10
- 1996年 Fujitsu VP2600/10E+VPP500/15+VX/2
- 1999年 Fujitsu VPP800/63
- 2004年 Fujitsu HPC2500
- 2008年 Fujitsu HX600（T2K）+M9000

詳しくはセンターパンフレットに記載があります
https://www.media.kyoto-u.ac.jp/accms_web/wp-content/uploads/2022/04/ACCMS2022-web.pdf

# Xeon Phi KNL導入前（T2K2世代目）

## 2012～2016年に稼働（20年来のF→Cへ）



# つい最近までの京大スパコン

## 2016～2022年（7/28に稼働停止）

### Camphor 2 (System A)

CRAY *XC40*

intel Xeon Phi KNL 68cores 1.4GHz x 1 /node

| | |
|---|---|
| #nodes | = 1,800 |
| #total cores | = 68 cores x 1,800 → 122,400 cores |
| Peak performance | = 3.05TFops x 1,800 → 5.48 PFlops |
| Memory capacity | = (96+16 GB) x 1,800 → 196.9 TB |
| Burst buffer | = 230 TB, 200 GB/sec DATAWARP |

### Storage

DataDirect NETWORKS *ExaScaler (SFA14K)*

| | |
|---|---|
| Disk capacity | = 24 PB |
| Bandwidth | = 150 GB/sec |
| Burst buffer | = 230 TB, 250 GB/sec |

IME

高速通信網 InfiniBand EDR/FDR

高速通信網 Omni-Path

### Laurel 2 (System B)

CRAY *CS400 2820XT*

intel Xeon Broadwell 18cores 2.1GHz x 2 /node

| | |
|---|---|
| #nodes | = 850 |
| #total cores | = 36 cores x 850 → 30,600 cores |
| Peak performance | = 1.21 TFlops x 850 → 1.03 PFlops |
| Memory capacity | = 128 GB x 850 → 106.3 TB |

### Cinnamon 2 (System C)

CRAY *CS400 4840X*

intel Xeon Haswell 18cores 2.3GHz x 4 /node

| | |
|---|---|
| #nodes | = 16 |
| #total cores | = 72 cores x 16 → 1,152 cores |
| Peak performance | = 2.65 TFlops x 16 → 42.4 TFlops |
| Memory capacity | = 3 TB x 16 → 48.0 TB |

京都大学 KYOTO UNIVERSITY

# 次期スパコンの構成検討

## 最初はこれまでの構成を継続したシステムを検討

2021年にリプレイスを想定

- システムAはメニーコア（90コア/ノード以上）＋高メモリバンド幅（HBM）マシン
- システムBは汎用的なx86マシン（DDR5でまあまあコア数も多め）
- システムCはBの一部を大容量メモリに（Optaneも考慮した16TB程度）
- 新しくGPUマシンも少し入れる（8-16ノード）
- ストレージはバーストバッファ的なのは無しでSSDとHDDの構成を
- インターコネクトはNDR世代
- AとBでそれぞれリプレイス前と同じノード数程度→ノード性能が大体2～3倍くらいとして、全体的に2倍以上の性能向上

# 想定されたハードウェアなど

## システムA向けCPUはARMを期待していた

ThunderX4やA64FX後継機などSC18、19頃の話しでは、512bitSIMD対応+HBMのメニーコアARM系CPUの選択肢があった。

- ThunderX2を購入して性能テストなど行っていた。

ところが、すべて実現せず・・・

- HBMが使えるCPUはA64FX、SX-AT、まだ見ぬXeonだけになった。
- ARM系CPU待ちでリプレイス時期を延期
- まだ見ぬXeon待ちで更にリプレイス時期を延期

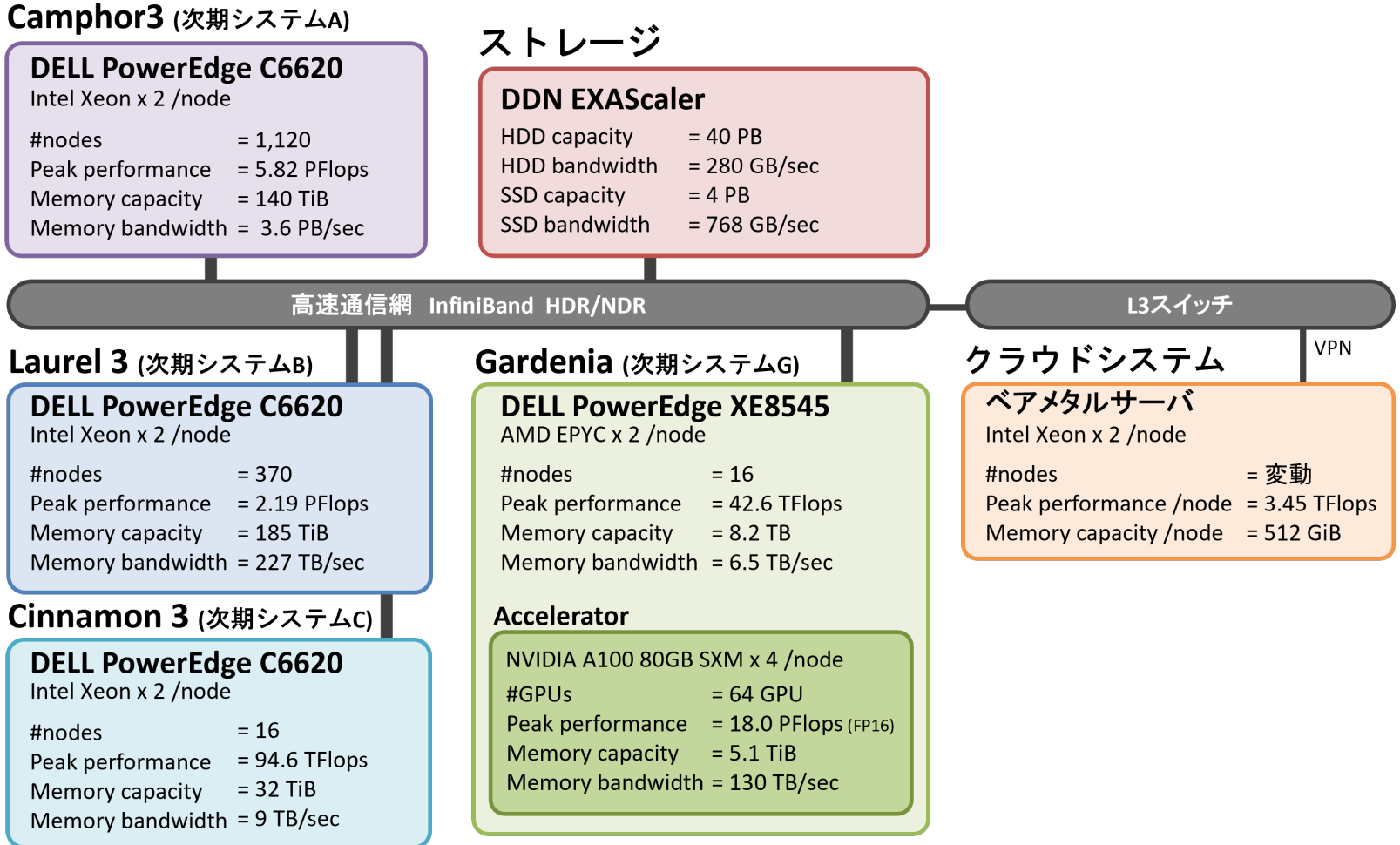
GPUはIntelのPVCも期待していたが・・・

- Auroraには導入されるはずなので、SC22で何か発表ある？

COVID-19とウクライナ危機で半導体不足と円安のダメージを受けて、想定していた物量が不可能に😱

# 次期京大スパコン構成

## 2022/11～（予定だったが延びそう）



# 次期京大スパコン詳細構成1

## システムA

Dell EMC PowerEdge C6620を1120ノード

- 最新XeonとHBMで高コア高演算性能高バンド幅のシステム

| ノード構成 | 仕様 |
|---|---|
| CPU | SPR Xeon (50コア以上, 2.5TF以上)×2 |
| メモリ | HBM2e (64GB, 1.6TB/s以上)×2 |
| 内蔵ストレージ | 480GB SSD |
| ネットワーク | InfiniBnad NDR (400Gbps) |
| 冷却 | 水冷 |
| OS | RHEL8 |

| システム構成 | |
|---|---|
| ノード数 | 1120 |
| 理論性能 | 5.8PF以上 |
| メモリサイズ | 143TB以上 |
| メモリバンド幅 | 3.5PB/s以上 |
| バイセクションバンド幅 | 18.6TB/s |

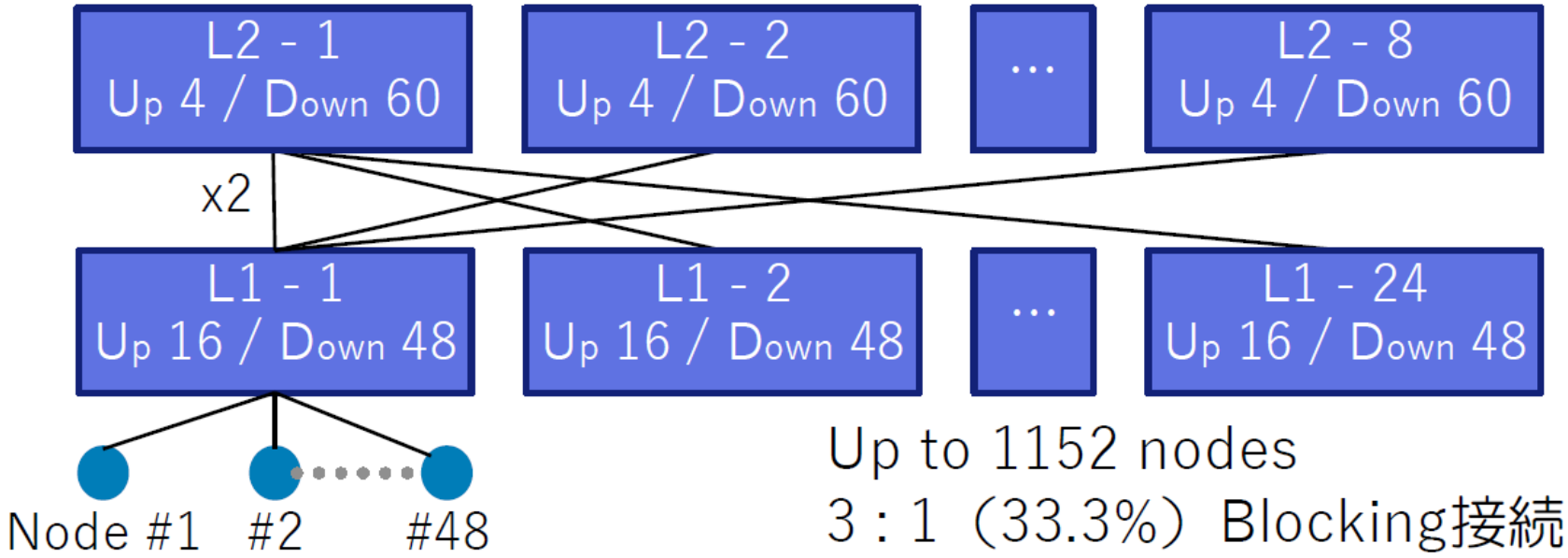
**Dell EMC PowerEdge C6620**
**x 1,120ノード（280シャーシ）**

# 次期京大スパコン詳細構成2

## システムA

Interconnect

- NVIDIA QM9790（NDR400Gbps）を2階層で構成するFat-tree。



# 次期京大スパコン詳細構成3

## システムBとC

Dell EMC PowerEdge C6620を370ノード(B)と16ノード(C)

- 最新XeonとDDR5で高コア高演算性能高メモリ容量のシステム

| ノード構成 | 仕様 |
|---|---|
| CPU | SPR Xeon (50コア以上, 5.2TF以上)×2 |
| メモリ | B: DDR5 (512GB, 600GB/s以上)<br>C: DDR5 (2TB, 550GB/s以上) |
| 内蔵ストレージ | 480GB SSD |
| ネットワーク | InfiniBnad NDR (200Gbps) |
| 冷却 | 水冷 |
| OS | RHEL8 |

| システム構成 | |
|---|---|
| ノード数 | 370 (B)+16 (B) |
| 理論性能 | 5.8PF以上 |
| メモリサイズ | 200TB以上 |
| メモリバンド幅 | 200TB/s以上 |
| バイセクションバンド幅 | 4.82TB/s |

Dell EMC PowerEdge C6620
x (370＋16)ノード(97シャーシ)

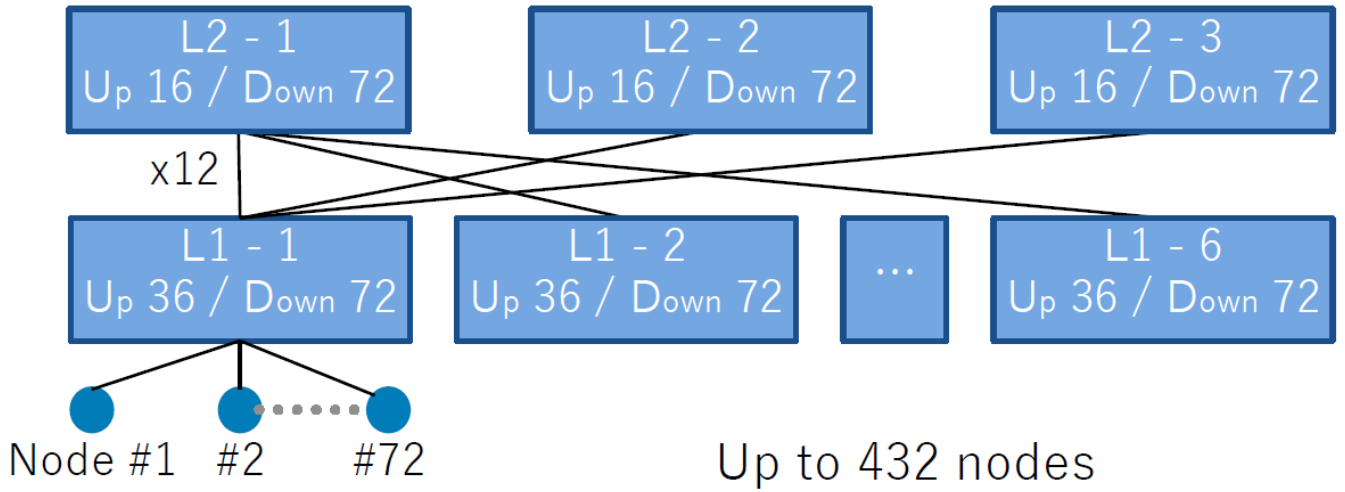
# 次期京大スパコン詳細構成4

## システムBとC

Interconnect

- NVIDIA QM9790を2階層で構成するFat-tree。ノード間は200Gbpsで、スイッチ間は400Gbpsで接続

サブシステムB Fat-tree NDR200 論理接続図




Up to 432 nodes
2 : 1 （50%） Blocking接続

# 次期京大スパコン詳細構成5

## システムG

Dell EMC PowerEdge XE8545を16ノード

- GPUを4枚/ノードのAI処理向けシステム

| ノード構成 | 仕様 |
|---|---|
| CPU | AMD EPYC 7513(32コア, 2.6GHz)×2 |
| メモリ | DDR4-3200 (512GB, 409GB/s) |
| GPU | NVIDIA A100 80GB SXM×4 |
| 内蔵ストレージ | 1.92TB SSD×2 |
| ネットワーク | InfiniBnad HDR (200Gbps) |
| 冷却 | 空冷 |
| OS | RHEL8 |

| システム構成 | |
|---|---|
| ノード数 | 16 |
| 理論性能 | 42TF+620TF |
| メモリサイズ | 8TB+5TB |
| メモリバンド幅 | 6.5TB/s+130TB/s |
| バイセクションバンド幅 | 800GB/s |

Dell EMC PowerEdge XE8545
x 16ノード（16筐体）

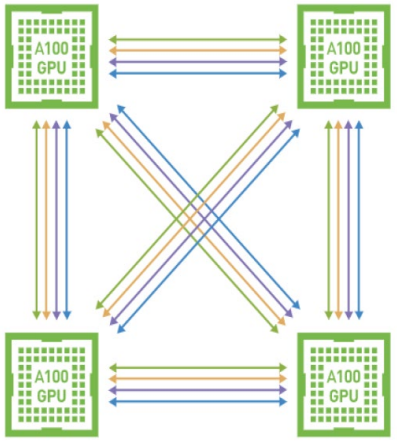
# 次期京大スパコン詳細構成6

## システムG

### GPU

- ノード当たり4機のA100を搭載し、NVIDIA NVLinkにより600GByte/秒の通信帯域での相互接続。

| | |
|---|---|
| 倍精度演算性能 | 9.7TF |
| 単精度演算性能 | 19.5TF |
| 半精度演算性能 | 312TF |
| メモリサイズ | 80GB |
| メモリバンド幅 | 2039GB/s |
| NVLink | 600GB/s |




### Interconnect

- NVIDIA QM9790を1階層で構成するFat-tree網

# 次期京大スパコン新規追加機能

## クラウドサブシステム

いわゆるクラウドバースティングを実装。

Oracle Cloud Infrastructure (OCI) BM.Optimized3.36 (Intel)

ノード数は需要によって決める予定（接続口用に1ノード準備）

11月下旬から利用開始予定

| ノード構成 | 仕様 |
| --- | --- |
| CPU | Intel Xeon Gold 6354 (18コア, 1.9TF)×2 |
| メモリ | DDR4 (512GB, 400GB/s?) |
| 内蔵ストレージ | 3.84TB NVMe SSD |
| ネットワーク | 2×50Gbps、1×100Gbps RDMA |
| OS | Oracle Linux 8 |

# 次期京大スパコンまとめ

## これまでと同様のシステムAとBとC＋GPUとクラウドを導入

システムA
- メニーコア、自作コードを持った研究者がプロダクトランに使うイメージ。

システムB
- Xeon+DDRの汎用計算機、ISVアプリやシングルノードユーザ向け。

システムC
- 大容量メモリ、共有メモリアプリ向け。

システムG
- GPUマシン、単ノード利用を想定（大規模並列GPUは他サイトで）。

クラウドシステム
- オンプレリソース不足、オンプレとは違う環境を使いたい人向け。

# VMサーバホスティングとの連携

## 新スパコンと10Gbps接続

センサデータを収集してスパコンで解析などの利用向け

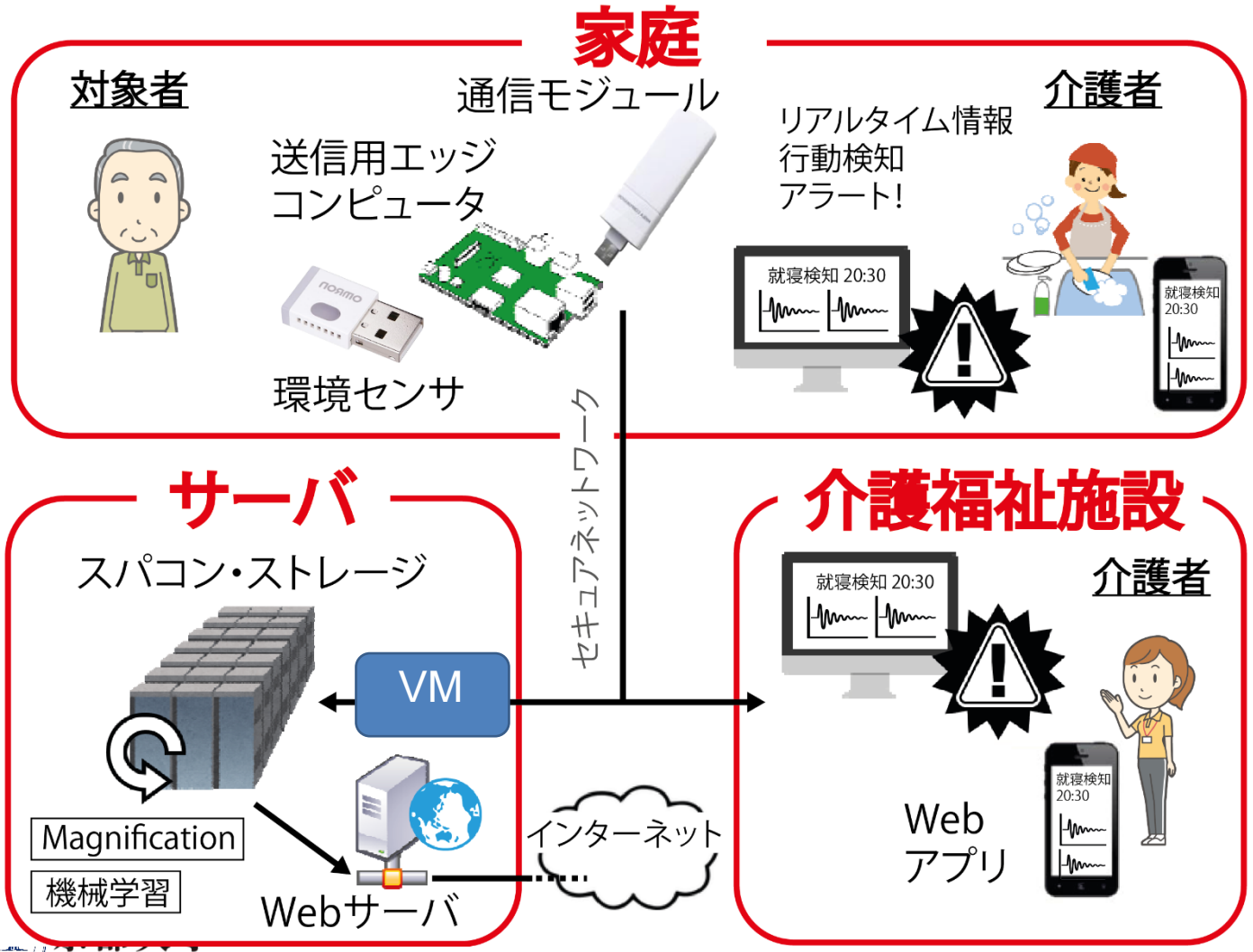
| | |
|---|---|
| 仮想化環境 | VMWare |
| OS | Alma Linux 8<br>RHEL 8<br>CentOS Stream 9<br>Ubuntu 20<br>Windows Server 2019 Datacenter |
| 標準スペック | CPU:2コア,メモリ:4GB,ディスク:100GB |
| SSHによるログイン | 可能(Root権限も付与) |

※CPUコア数・メモリ容量・ディスク容量は増量可能。RHELは別途費用がかかる

| | 利用負担金 |
|---|---|
| VMサーバ | 12,000円/年（京大内基本料金） |
| システム資源増量 | CPU:2コア単位で3,000円/年<br>メモリ:4GB単位で3,000円/年 |
| ディスク増量 | 100GB単位で6,000円/年 |

# VMとスパコンを利用した研究例

## IoT機器を使った見守りシステム



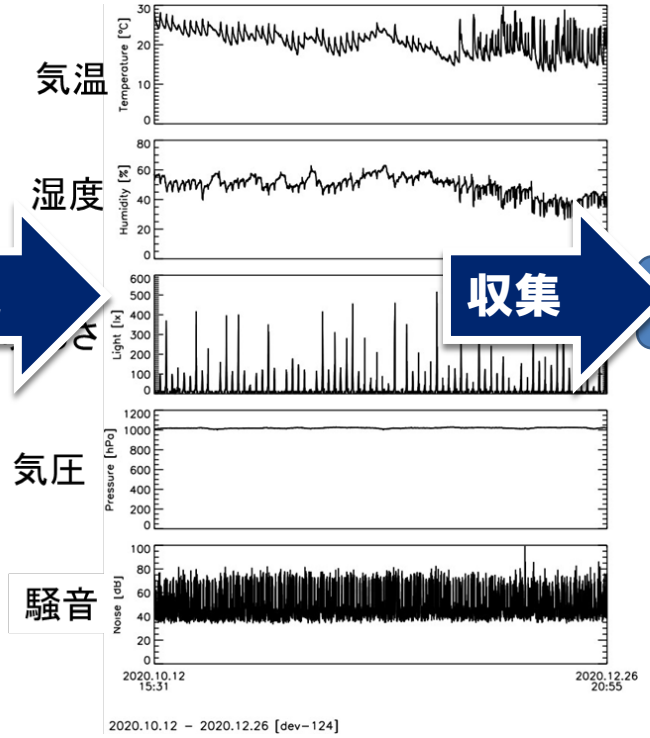

数値シミュレーションだけでなく、新しいスパコンの利用方法へのチャレンジ。

センサに繋がったIoT機器から、データがVMからスパコンに送られて、スパコンで解析し、検知、予測などを行うシステム。

# VMとスパコンを利用した研究例

## IoT機器からVM経由で収集されたデータを使いスパコンで処理

測定

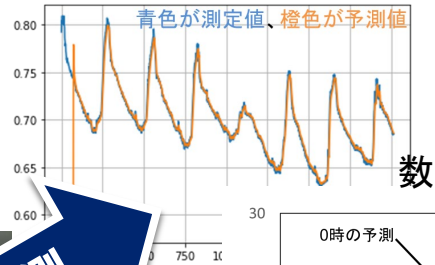



収集

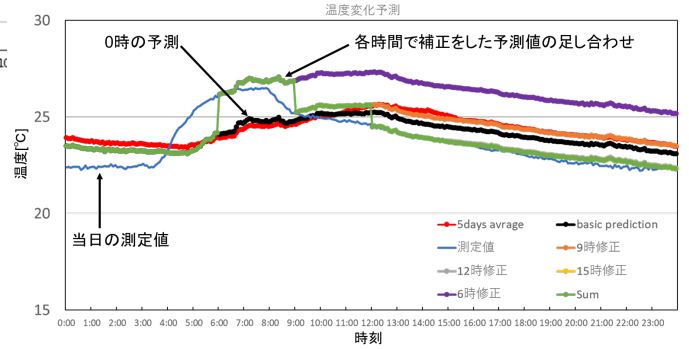
VM


予測

RNNによる気温予測

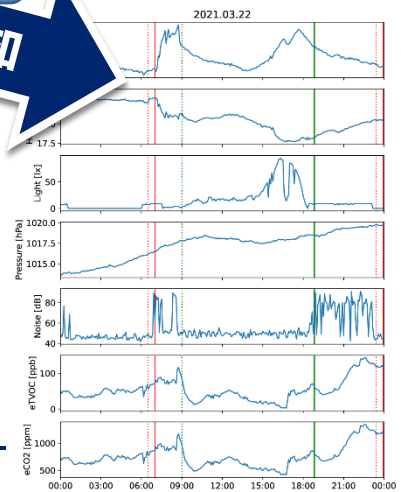



数値モデルによる気温予測




VM

検知

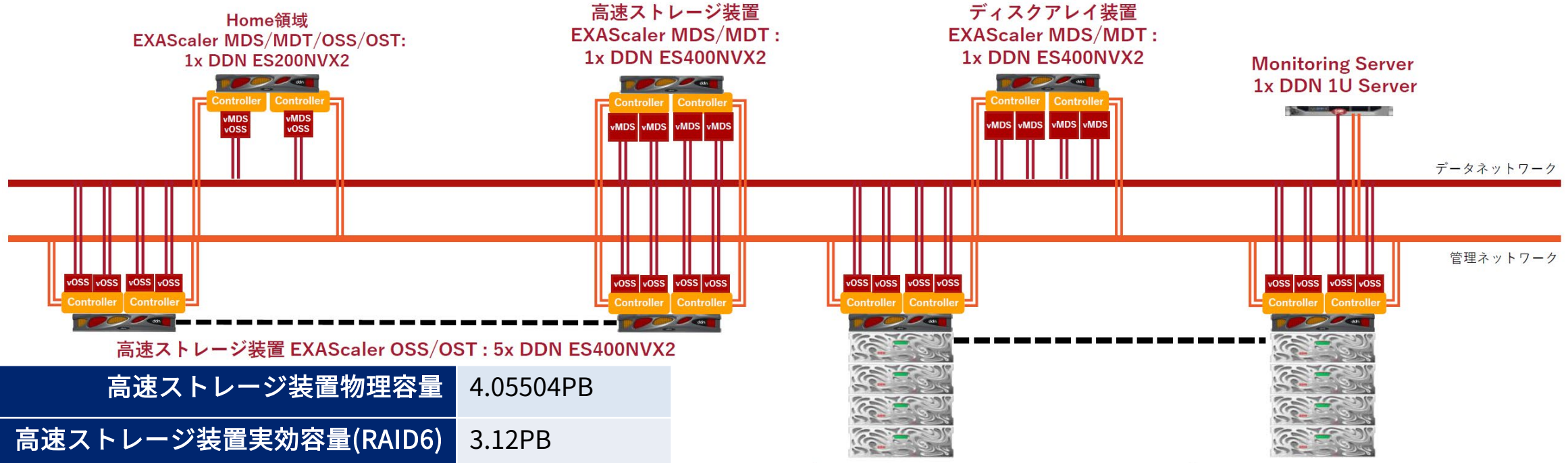



起床：○，外出：○，就寝：○

見守りの負担軽減

# 次期京大スパコンストレージ

## 高速＋大容量のストレージ構成




| | |
|---|---|
| 高速ストレージ装置物理容量 | 4.05504PB |
| 高速ストレージ装置実効容量(RAID6) | 3.12PB |
| 高速ストレージ装置データ転送速度 | 768GByte/秒 |
| ディスクアレイ装置物理容量 | 40.32PB |
| ディスクアレイ装置実効容量(RAID6) | 31.99PB |
| ディスクアレイ装置データ転送速度 | 280GByte/秒 |

旧ストレージから
- 総データ量 11,839,27 TB
- 総ファイル数 4,824,242,200

を新ストレージへ1ヶ月で移行

# スケジューラ

## 京大独自仕様のスケジューラ機能

京大では、占有、優先、準優先というサービスをしており、その運用上スケジューラに独自機能が必要。

※詳しくは8ページにあるジョブ管理ソフトウェアに関する要求要件を読んでください。

前回はPBSだったが、色々問題があり今回はSLURM（Simple Linux Utility for Resource Management）に変更。

- オープンソースのジョブスケジューラとして 2002 年にローレンス・リバモア研究所(LLNL)で開発
- SchedMD社が有償サポート(L3レベル)を提供

SLURMでもカスタマイズをすると、京大仕様を満たすことができる。

→スケジューラ縛りから開放

# まとめ

## 京大スパコンは絶賛（色んな意味を含む）リプレイス中

- ✓ もともとは2021年頃にリプレイス予定だったが、色々起き、現在リプレイスが進んでいる。
- ✓ これまでの構成を踏襲しつつ、GPUやクラウド機能も導入している。
- ✓ COVID-19やウクライナ危機で、元の想定した旧システムからの2-3倍の性能向上は難しくなったが、実利用では高い性能を期待している。
- ✓ 京大スケジューラ要件は、縛りが高いといわれることがあったが、大体どこでも実装可能と分かり、自由度が上がった。
- ✓ ISVアプリケーションはなぜか激値上がりしており、今後も取捨選択が続きそう。
- ✓ 次期システムはどうなるか分からないが、PVC次第な気がする。