# Zabbixを用いたGPU基盤の監視について

トヨタ自動車 株式会社

# 自己紹介

- 名前：梅津 拓海（うめつ たくみ）
- 所属：トヨタ自動車　デジタル情報通信本部　InfoTech情報通信先行開発室
- 業務内容：7年間インフラエンジニアとして業務に従事。2024年にGPU基盤導入に初めて携わる。Zabbix初心者なのでお手柔らかに…

# アジェンダ

- Zabbix導入の理由
  - 交通事故ゼロ社会に向けて -GPU基盤の導入-
  - GPU基盤に必要な監視
  - GPU基盤の監視 –Zabbixが担う役割-
  - GPU故障監視の必要性
- GPU基盤の運用
  - 機器構成
  - 実際に運用してみて
  - NVIDIA GPU監視テンプレートについて
- 今後の展望

# アジェンダ

- Zabbix導入の理由
  - 交通事故ゼロ社会に向けて -GPU基盤の導入-
  - GPU基盤に必要な監視
  - GPU基盤の監視 –Zabbixが担う役割-
  - GPU故障監視の必要性
- GPU基盤の運用
  - 機器構成
  - 実際に運用してみて
  - NVIDIA GPU監視テンプレートについて
- 今後の展望

# 交通事故ゼロ社会に向けて -GPU基盤の導入-

- 「交通事故ゼロ社会」に向けた取り組みの一環として、GPU基盤を導入
- 主な利用用途：高度な自動運転システム、および危険を予知するAIエージェント等の開発

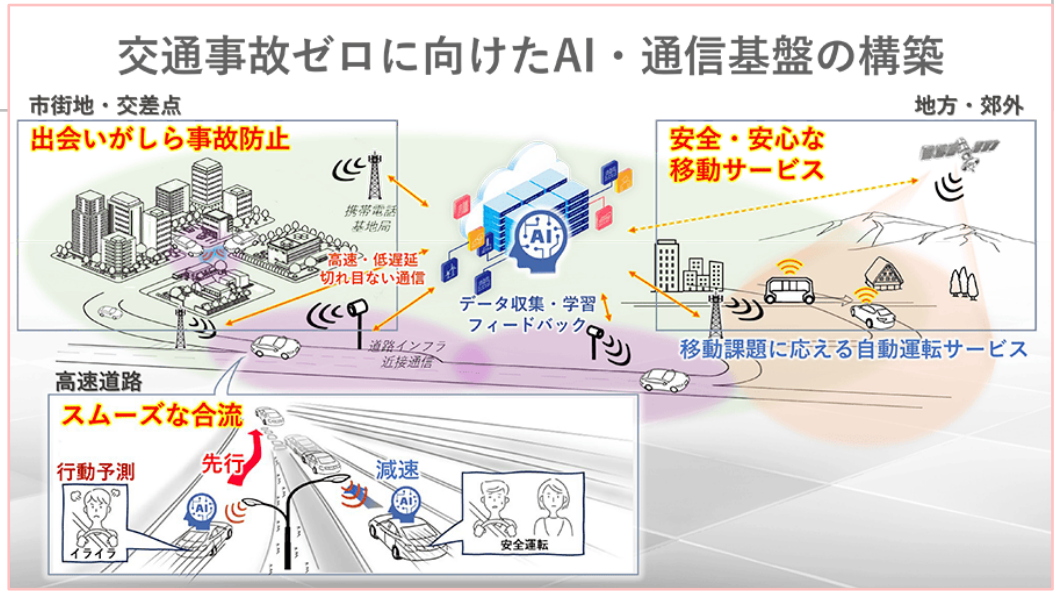
**NTTとトヨタ自動車、交通事故ゼロ社会の実現に向けた「モビリティ×AI・通信」の共同取り組みに合意**

https://global.toyota/jp/newsroom/corporate/41746612.html

利用用途

自動運転システムの開発

AIエージェントの開発



# GPU基盤に必要な監視

- 通常のITインフラシステムと比較して、GPU基盤の監視にはGPUリソースの監視などが必要となる
- 最初からすべてを導入するのでなく、スモールスタートで死活監視/リソース監視ができる監視基盤を構築する方針で進めていた

| 監視項目 | 監視内容 |
|---|---|
| 設備監視 | 電源供給状態、フロア/ラック内温度、ラック内の結露・漏水状況等を監視する |
| **死活監視** | 各機器の稼働状況等を監視する |
| **リソース監視** | GPUなどシステムリソースの利用状況を監視する |
| ネットワーク監視 | トラフィック量、帯域使用率、パケットロス、レイテンシ等を監視する |
| ログ監視 | システム・アプリケーションのログを収集/解析し、エラーや異常がないか監視する |
| ジョブ監視 | 実行したジョブが正常に終了しているか等を監視する |

今回はここの話

# GPU基盤の監視 –Zabbixが担う役割-

- トヨタGPU基盤においてZabbixが担う役割は「死活監視」
- Why Zabbix？ → オンプレミス基盤の死活監視に多くの実績があり安定したプロダクトのため
- 各ツールの特徴にあわせた役割分担を行いながら、DC全体の監視を実施する

| | Zabbix | Prometheus&Grafana |
|---|---|---|
| 主な役割 | **死活監視** | リソース監視 |
| 監視対象 | ITインフラ全般 | GPUサーバ、クラスタ管理ノード |
| 監視方法 | ping、SNMP、Zabbix agent | NVIDIA dcgm-exporter 等 |
| 監視項目 | ネットワーク疎通監視、ログ監視 | GPU利用率 等 |

# GPU故障監視の必要性

- 今回はNVIDIA GPU + Zabbixというトヨタでも初の試み（通常の監視 + α が必要）
- GPUが故障している時間が長いとAIモデル学習効率が悪くなる
- 常に高い負荷がかかるGPUを監視することは特に重要

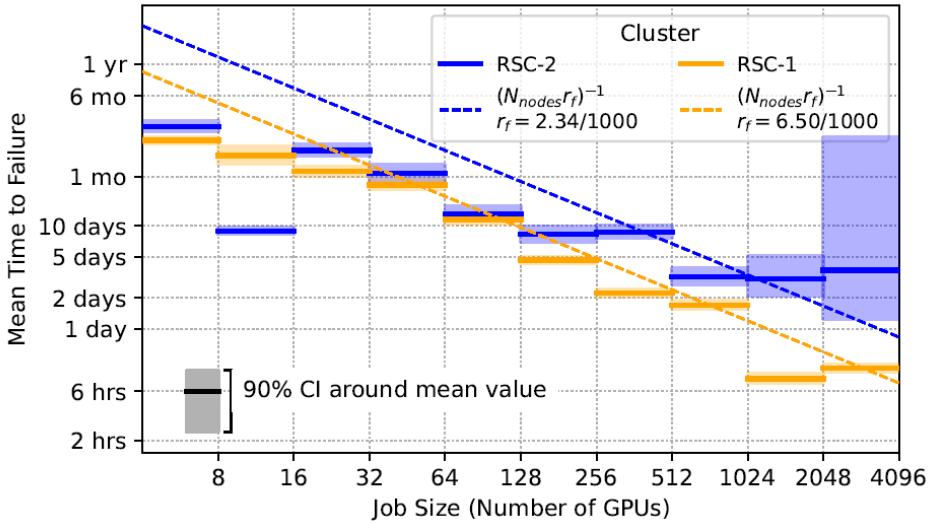
【Metaの論文より引用】

”1つのジョブで利用するGPUを増やした場合のMTTF（平均故障間隔）”

**13万GPUあると約10分に1回GPUが壊れるという試算**

- 8 GPUジョブ: 約47.7日
- 1024 GPUジョブ: 約7.9時間
- 16384 GPUジョブ: 約1.8時間
- 131072 GPUジョブ: 約0.23時間

引用元 https://ai.meta.com/research/publications/revisiting-reliability-in-large-scale-machine-learning-research-clusters/

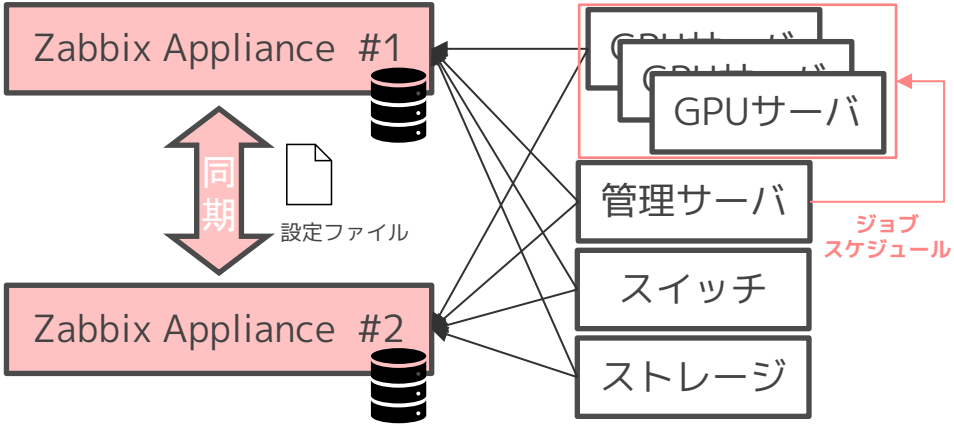


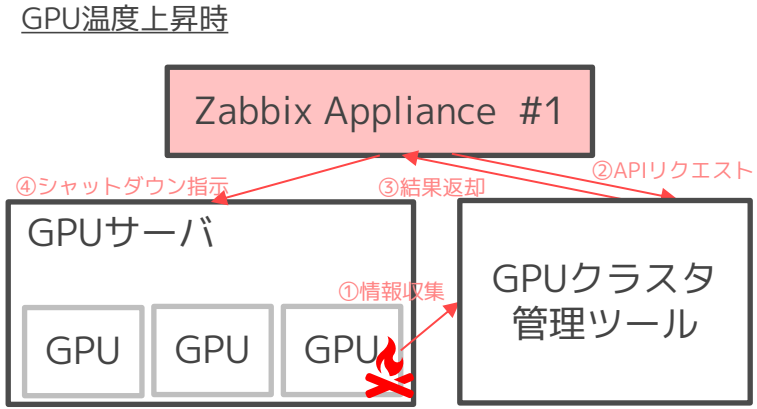
# アジェンダ

- Zabbix導入の理由
  - 交通事故ゼロ社会に向けて -GPU基盤の導入-
  - GPU基盤に必要な監視
  - GPU基盤の監視 –Zabbixが担う役割-
  - GPU故障監視の必要性
- GPU基盤の運用
  - 機器構成
  - 実際に運用してみて
  - NVIDIA GPU監視テンプレートについて
- 今後の展望

# 機器構成

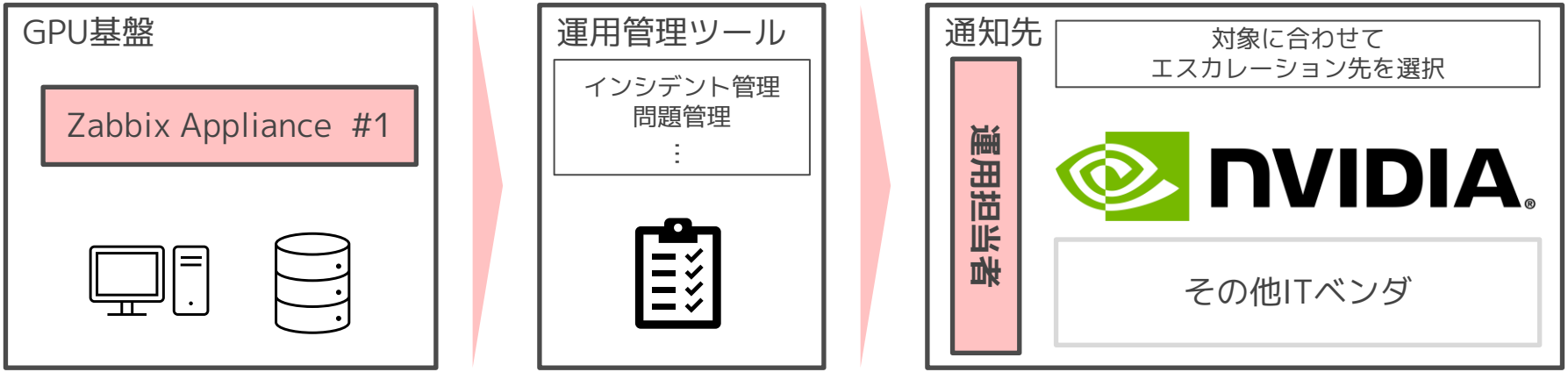
- 実際のZabbix機器構成として、Zabbix Appliance x2台で冗長構成を組んでいる
- 温度が閾値を超えた場合に対象GPUサーバをシャットダウンする仕組みをスクリプトで実装
  - ① GPUクラスタ管理ツールがGPUサーバからGPU温度を定期的に取得
  - ② Zabbix Applianceがクラスタ管理ツールの**APIを定期的にたたいてGPU温度を監視**
  - ③ APIで返却されたGPU温度が閾値を超えて上昇していることを検知
  - ④ GPUサーバOSに対して**シャットダウン指令**を出す



# 実際に運用してみて (1/2)

- 運用の流れ：GPU基盤/Zabbix → 運用管理ツール → 各ITベンダのサポート
- 運用を開始してからよく発生するエラーは下記
  - ① NVIDIAの物理機器関連：GPUが壊れた…
  - ② ジョブスケジューラ関連：ジョブがfailした…
  - ③ ファイルシステム関連：クライアント側サービスの起動に失敗した…



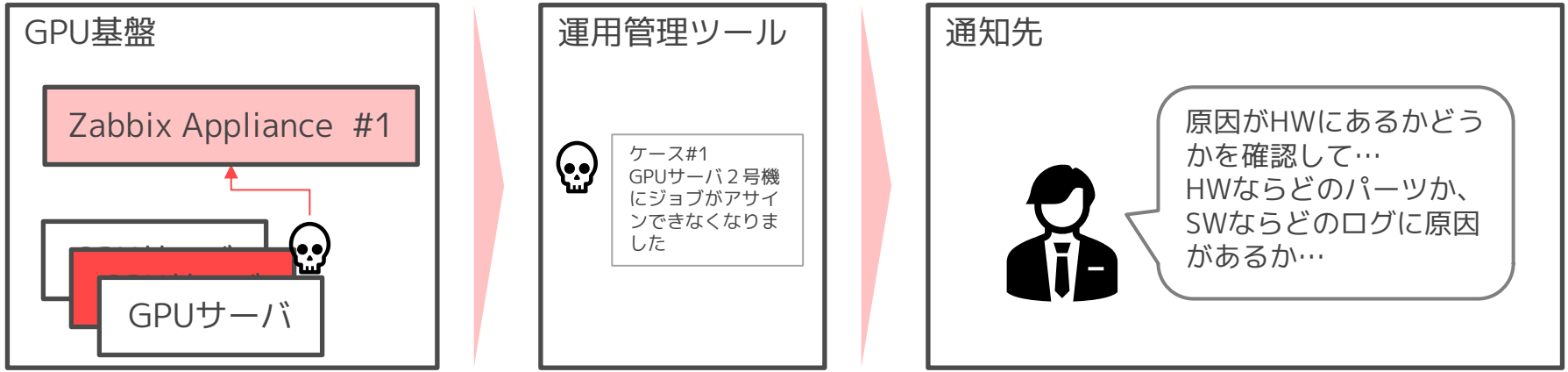
# 実際に運用してみて (2/2)

- Zabbixでは当初GPUサーバのシステムログ（認証ログ、カーネルログ等）のみ監視していたが、それのみだとジョブスケジューラ上のステータスを確認できなかった
- 特に何らかの理由でジョブのアサインができなくなったGPUサーバを検知・通知したいため、システムログ以外にジョブスケジューラのログを監視するよう設定

① GPUサーバが故障したことをログから検知

② インシデントとして起票

③ 運用担当者にて現状復帰/原因追及



# NVIDIA GPU監視テンプレートについて (1/4)

- 2024年11月に、Zabbix 7.2でNVIDIA GPUをモニタリングするためのテンプレートおよびZabbix agent2用プラグインがリリースされた
- ただしこの時点ではコンシューマ向けモデル(NVIDIA GTX 1650s、NVIDIA RTX 2070Ti)のみ対応
- データセンター向けGPU(Hopper)でも利用できるかは未知数な状況であったが、GPU監視の手段は多い方が良いので試してみることにした

Zabbix + NVIDIA

● **NVIDIA GPU monitoring template and Zabbix agent 2 plugin**

https://blog.zabbix.com/see-whats-possible-in-zabbix-7-2/29373/#NVIDIA_GPU_monitoring_template_and_Zabbix_agent_2_plugin

GPU monitoring

Tested versions

This template has been tested on:

- Nvidia GTX 1650s
- Nvidia RTX 2070Ti

**Zabbix社にてテスト済み**


NVIDIA RTX 2070Ti

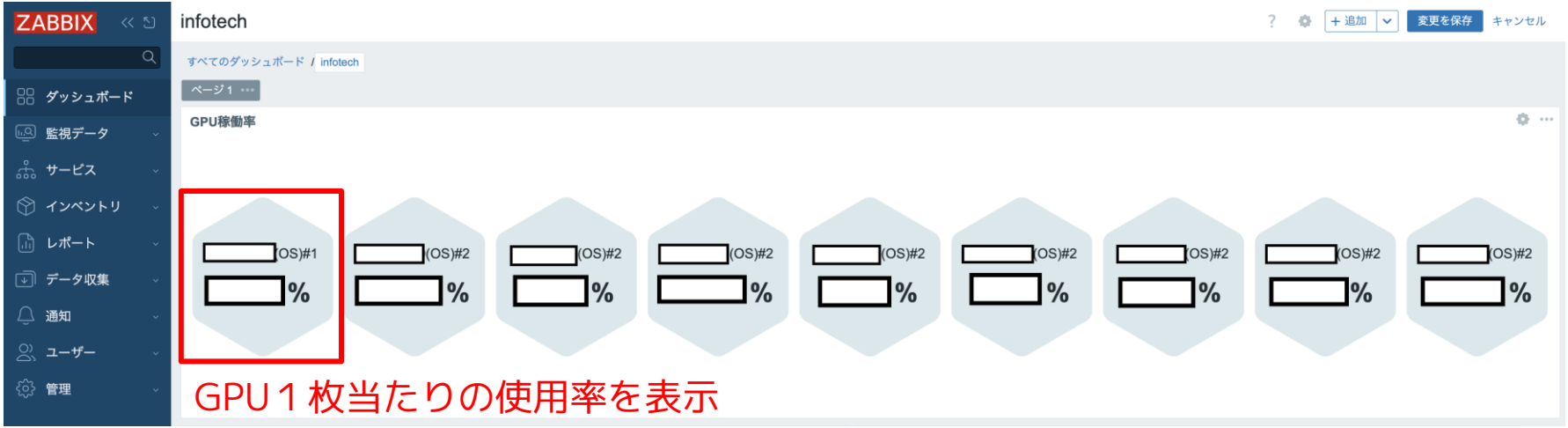
**Zabbix社にてテスト未実施**

NVIDIA H200

画像はNVIDIA公式より引用

# NVIDIA GPU監視テンプレートについて (2/4)

- データセンター向けGPUからもメトリックを取得できた
- 一部データセンター向けGPUには存在しないFANの監視等が含まれていたが、プロトタイプを無効化して対応
- テンプレート追加のみで、簡単にGPU1枚ごとのメトリックを取得

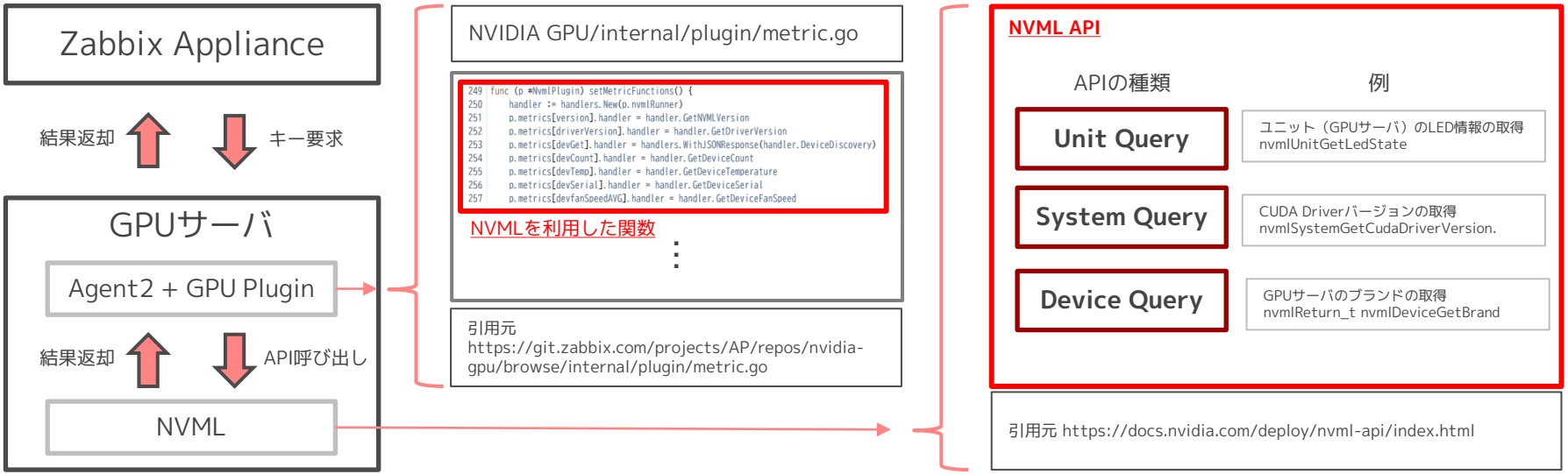


# NVIDIA GPU監視テンプレートについて (3/4)

- GPUデータの取得手段として、裏ではNVML APIを利用している

  ※NVML（NVIDIA Management Library）：NVIDIA GPUのモニタリング用ライブラリ

- NVMLを利用したり、NVMLと似たDCGM-ExporterをWrapしてZabbixにメトリックを送信するなどして色々なことができそうだと感じた

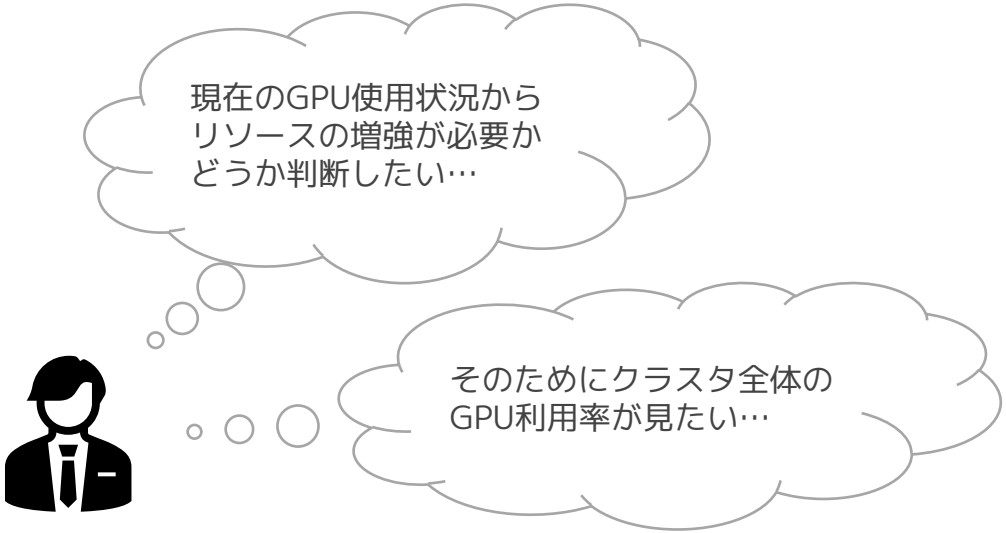


# NVIDIA GPU監視テンプレートについて (4/4)

- GPUがUUIDで表示されるため、そのままでは使いにくい
- デフォルトで、UUIDでないかたちでわかりやすく表示されると嬉しい（GPU1、GPU2…など）
- またデフォルトで個々のGPUでなくGPUサーバ単位、クラスタ単位でGPUメトリックを確認できると嬉しいと感じた

【現状のデフォルトの表示】GPUサーバ1台の表示例

| | |
|---|---|
| [GPU-2f7c6c51-3c4a〜〜]: GPU utilization | グラフ |
| [GPU-31d6dbc1-2d94〜〜]: GPU utilization | グラフ |
| [GPU-7f291e25-c325〜〜]: GPU utilization | グラフ |
| [GPU-b54d2fe5-22d5〜〜]: GPU utilization | グラフ |
| [GPU-abb7ce4d-012c〜〜]: GPU utilization | グラフ |
| [GPU-cd19a73b-eb1c〜〜]: GPU utilization | グラフ |
| [GPU-f7c1c062-9c13〜〜]: GPU utilization | グラフ |
| [GPU-a062cfbb-06f1〜〜]: GPU utilization | グラフ |



# アジェンダ

- Zabbix導入の理由
  - 交通事故ゼロ社会に向けて -GPU基盤の導入-
  - GPU基盤に必要な監視
  - GPU基盤の監視 –Zabbixが担う役割-
  - GPU故障監視の必要性
- GPU基盤の運用
  - 機器構成
  - 実際に運用してみて
  - NVIDIA GPU監視テンプレートについて
- 今後の展望

# 今後の展望

- トヨタGPU基盤におけるZabbixの役割としては「死活監視」
- 引き続きZabbixは死活監視に特化させつつ、リソース監視はPrometheus&Grafanaで行う
- データの可視化はGrafanaで行う（Zabbix plugin for Grafanaを利用）

**死活監視**

ZABBIX

**リソース監視**

Prometheus

Zabbix plugin for Grafana

**監視インターフェース**

Grafana