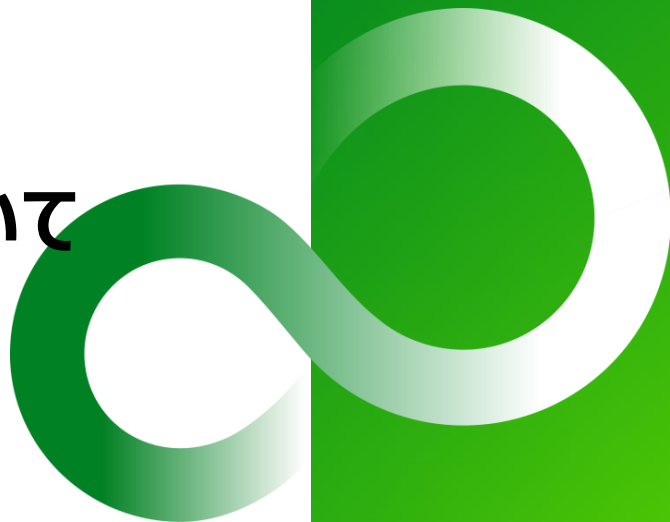


# KubeCon EU 2024登壇について

2024年04月23日

富士通株式会社 先端技術開発本部

園田雅崇



# 本日も話す内容

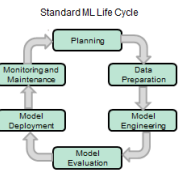


- 発表した内容の簡単な紹介
- 登壇して感じたことなど
  - 所感
  - 登壇して良かったこと
  - 採択のために意識したこと

# 発表概要

## Accelerators(FPGA/GPU) Chaining to Efficiently Handle Large AI/ML Workloads in K8s

# 背景、課題

- 背景(やりたいこと)：k8s/アクセラレータ(Acc)を活用してAI/MLを高効率・高性能に実施したい
  - k8sがバックエンドで動くAI/MLフレームワークが増えている
  - 大規模なAI/MLワークロードには汎用CPUよりもAccを活用した方が効果的
  - 現状のk8sではAcc活用は限定的(Podと紐づいており、各処理間通信はPod間通信となりCPU介在)
- 課題：今のk8sではアクセラレータを効果的に活用出来ない
  - 非コンポーザブル：アクセラレータ間の接続及び高速化に必要な設定は手動なので、構成が固定され構成変更は困難。
  - CPUオーバーヘッド：アクセラレータ間で通信を行う際にCPUが介在してホストにおけるメモリコピーなどを行うため

Kubernetes: Empowering AI&ML DevOps	Acceleration Methods and Advanced Setups	Challenges of Process Acceleration in K8s
<p>In recent years, with the widespread adoption of AI/ML, AI/ML development frameworks backed by Kubernetes (K8s) have become increasingly utilized.</p>  	<ul style="list-style-type: none"> <li>• <b>Acceleration methods in Kubernetes (K8s):</b> <ul style="list-style-type: none"> <li>○ Accelerators are allocated to Pods processing each step of a pipeline.</li> <li>○ Processing is offloaded to accelerators from within the Pods.</li> </ul> </li> <li>• <b>For further acceleration through efficient data communication between accelerators:</b> <ul style="list-style-type: none"> <li>○ Implement SR-IOV/RDMA settings</li> <li>○ Establish communication paths through device NICs and configure their network settings.</li> <li>○ Connect components via PCIe.</li> <li>○ Facilitate data transfer through host memory.</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• <b>Non-Composable</b> <ul style="list-style-type: none"> <li>○ Fixed and hard-to-change system configuration</li> <li>○ Difficult system modifications in response to changing requirements.</li> <li>○ Inability to reuse and effectively utilize resources.</li> </ul> </li> <li>• <b>Low customization and integration capabilities</b> <ul style="list-style-type: none"> <li>○ Difficulty managing accelerated system parts with Kubernetes, complicating integration with other systems, or parameter adjustments.</li> </ul> </li> <li>• <b>Vendor lock-in</b> <ul style="list-style-type: none"> <li>○ Dependency on a single vendor, restricting flexibility and innovation potential.</li> </ul> </li> <li>• <b>Cost-effectiveness</b> <ul style="list-style-type: none"> <li>○ High costs due to the inability to adapt to scale changes and the need to allocate redundant resources in advance.</li> <li>○ High operation costs due to the prevalence of custom parts and application-specific aspects.</li> </ul> </li> <li>• <b>CPU overhead limiting performance</b> <ul style="list-style-type: none"> <li>○ Increased overhead from CPU processing tasks due to CPU intervention in communication between accelerators, such as memory copying in the host.</li> </ul> </li> </ul> 

# 我々のアプローチ：アーキ概要

- 我々のアプローチ：k8sの拡張モデルであるCR(CustomResource)の活用
  - CRを用いてk8sのリソースモデルを拡張することで、AccやAcc間のコネクションをk8s nativeなリソースとして管理
  - リソースモデル：Pipeline Definition, Abstracts, Individual Functions/Connectionsの3層
  - オペレータ：上記3層+スケジューラ

## Architecture of K8s Native Accelerator Chaining



Kubernetes Custom Resource as an extension model to manage accelerators and their connections

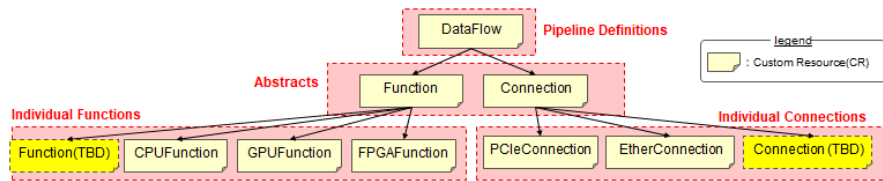
### Resource model for K8s native resource management:

Pipeline Definitions: Accelerator chain orchestration and development framework integration

Abstracts: Provide abstract layer to Manage accelerator and connections resources.

Individual Functions/Connection: Manage hardware resources for functions and connections

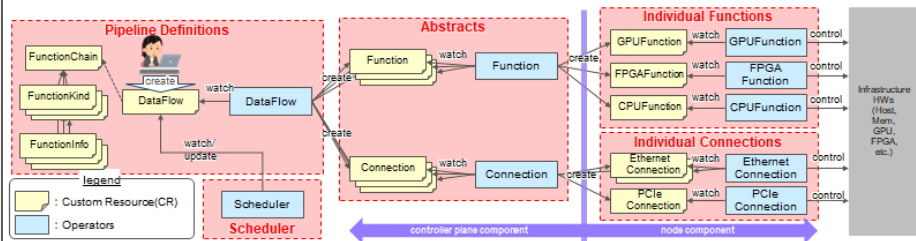
Provide K8s APIs to deploy, configure, and manage the lifecycle of accelerators and the connections between them.



## Operators Overview and Flow of Deployment



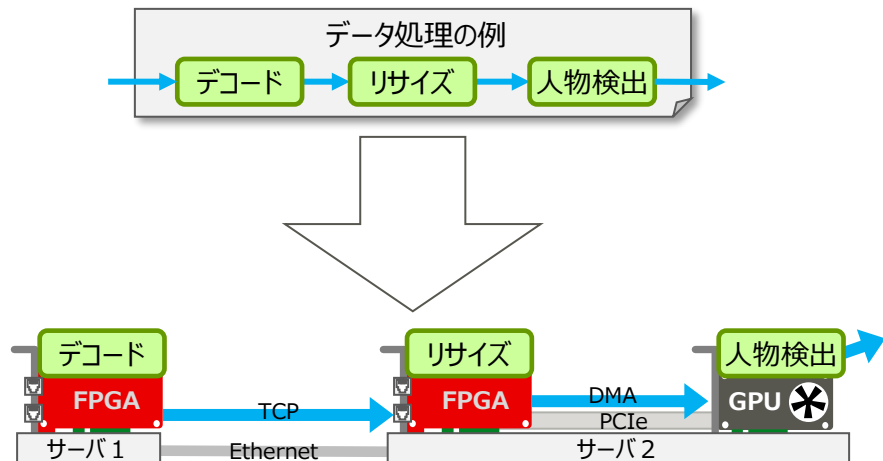
- **Pipeline Definitions:** Enable the creation of pipelines in composable way
- **Abstracts:** Create individual Function/Connection CRs according to accelerator(device)/connection type
- **Individual Functions:** Perform our own control for each type of accelerator(device)
- **Individual Connections:** Configure unique settings to both accelerators(devices) at end of the network path
- **Scheduler:** Determine where pipelines are deployed for performance and power efficiency



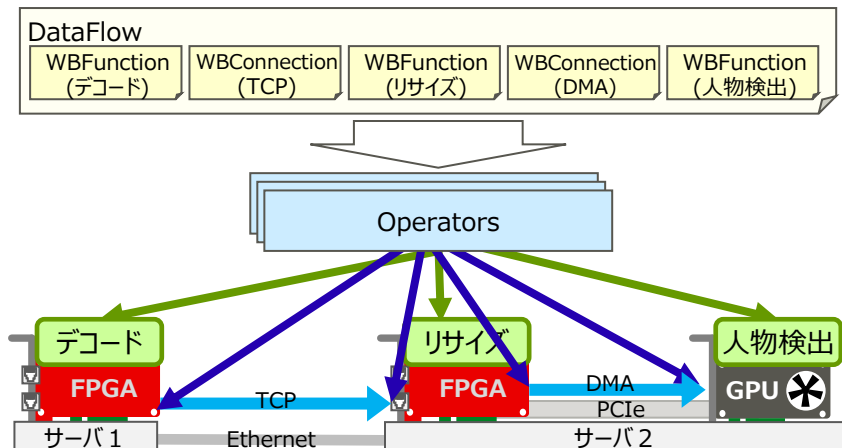
# 補足：具体例

- 映像解析(人物検出)のパイプラインをDataFlowで定義
  - 各step(デコード, リサイズ, 人物検出)をWBFunction CRとして定義
  - step間の通信手段をWBConnection CRとして定義
- 各CRに対応するOperatorが配備を実施

## やりたいことの例



## CRs & Operatorsで実現

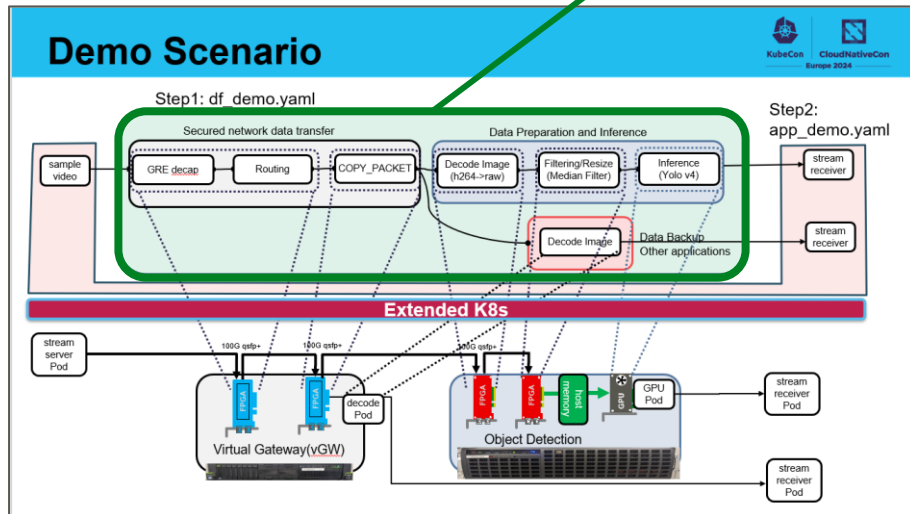
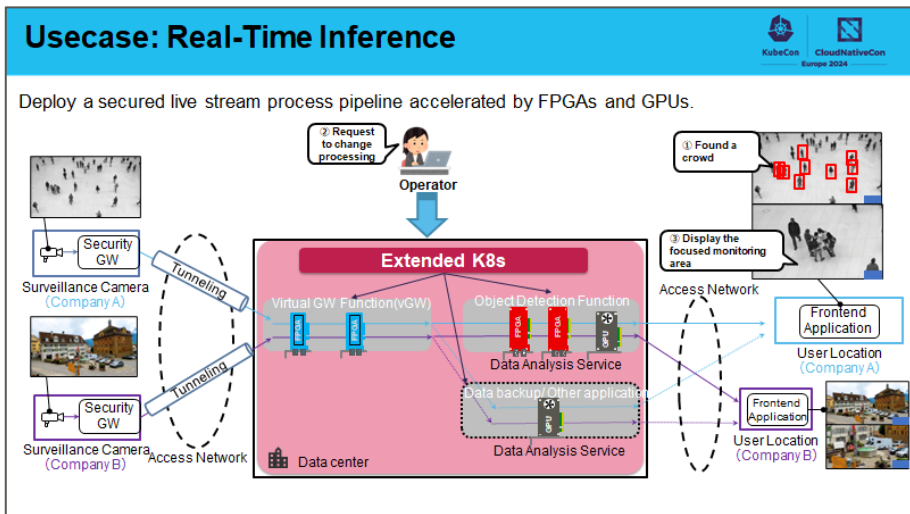


○映像解析のパイプラインを”DataFlow”として配備して以下を示した

- ”DataFlow”で定義された各種CRが作成される様子を表示
- 実際に映像を流して解析結果(BB付き映像)を表示

DataFlowの処理内容

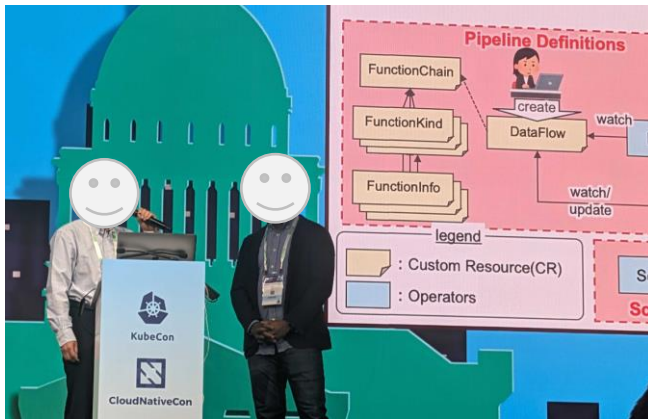
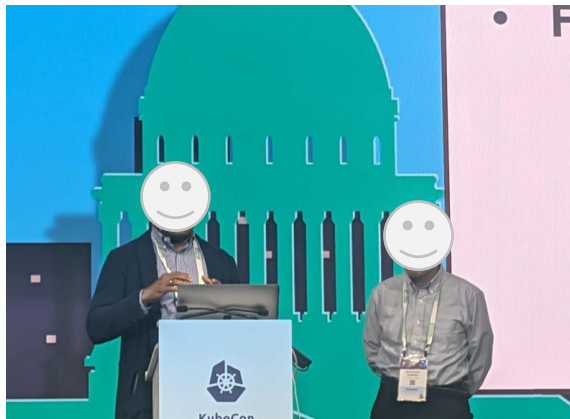
- ・セキュアネットワークへの転送処理
- ・映像解析処理：  
画像のデコード⇒フィルタリング&リサイズ⇒推論



# 登壇して感じたこと



- 過去に自身が参加したアカデミックな国際会議や展示会(MWC)とは別種の熱気があり、新鮮な印象
  - プレゼン(セッション)外での参加者同士の技術議論が活発。
  - プレゼンだけでなく、展示会的なエリア(Solutions SHOWCASE : 後述)もあり、こちらはお祭りのような雰囲気も感じた
- 自分達の発表に関しても、セッション後に声をかけられることもあった。
  - 今までに無い経験。それだけ興味を持ってくれたのか、KubeConが交流が盛んな文化なのか。
- 普段の業務ではなかなか各プロジェクトの取組みをチェック出来ていなかったもので、様々なプロジェクトや取組みを知れる良い機会になった。



# KubeConに登壇して良かったこと

- 自分達の取組みを広く(全世界に)広報出来る
  - CloudNativeに関する世界最大規模のイベントなので注目度も桁違い
- 他のWGの方と直接会話ができる
  - 我々のプレゼンを聞いてくれたら話が早い
  - どこかのコミュニティに持っていきたい場合などは、プレゼンがきっかけとして受け入れられやすくなる？
- 英語力向上の機会になる
  - 発表スライドやオーラル作成のためのWriting、質疑応答向けのListening, Speakingなど
  - 登壇までの間は超短期とはいえ集中トレーニングして対応した
- 関連するコミュニティに声をかけてもらえた

## ○トレンドに即したストーリーで考える

- 今回で言えば、AIに活用するという流れで発表内容を整えていった
- 実際今回のKubeConはAIをかなり前面に出していて、AI関連のセッションも多かったので、採択につながった一因では無いかと思う

## ○何度かトライを続ける

- 我々の場合は、KubeCon NA 2023にもエントリー申請していて2回目のエントリーで採択された
- タイミングもあるかと思うが、繰り返しエントリーを続けることも大事かも

**Thank you**

