INSPUR FPGAカード

2020/05

Inspur FPGA アクセレレーションカード

特殊

汎用

Inspur製

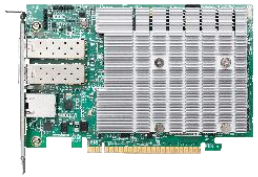
既　製

F07V

Media Codec + AI

F07V, ZU7EV *2
16 Streams Codec, FHHL

N10X



SmartNIC

N10X, Xilinx ZU17EG+CX4L,
2*25GE, FHHL

F10A

Intel GX1150, 1.366TFLOPS,
HHHL

F10S

Intel GX2800, 9.2TFLOPS,
FHHL

F37X

Xilinx VU37P, 28.1TOPS, HBM2
8GB, FHHL

PAC

Intel FPGA, GX1150,
1.366TFLOPS, HHHL

U200

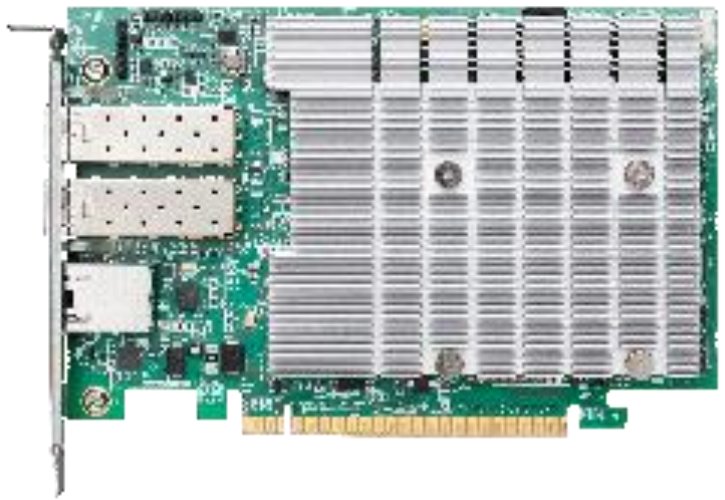
Xilinx FPGA, VU9P(XCU200),
18.6TOPS, FH3/4L

U250

Xilinx FPGA, VU13P(XCU250),
33.3TOPS, FH3/4L

inspur

N10X: ネットワークとコンピューティングを加速するスマートNIC



主要コンポーネント	
サイズ	FHHL Single-Slot
FPGAチップ	XCZU17EG-FFVC1760: quad-core ARM® Cortex-A53 up to 1.5GHz, dual-core Cortex-R5 real-time processors up to 600MHz, 926 K System Logic Cells, and 1590 DSP Slices
NICチップ	Mellanox ConnectX-4 Lx
ストレージ	
FPGA PS side	8GB DDR4(64bit DATA+ 8bit ECC) @2400MTs on-chip memory
FPGA PL side	4GB DDR4(64bit DATA+ 8bit ECC) @2400MTs on-chip memory
その他	16MB SPI Nor FLASH; 1Gb*2 QSPI Nor FLASH; 512Kb EEPROM
インターフェイス	
システム	PCIe Gen3 x16: FPGA PCIe x8 + CX4 PCIe x8
ネットワーク	2x 25GbE SFP28
テスト	3-line UART Test Interface（LVTTL level）
その他	Tri-speed RJ45 管理ポート; 6pin JTAGポート
電力 & 放熱	
電　源	PCIe-slot 12V@75W
消費電力	45W (Peak)
冷　却	放熱板冷却
ボード管理	
BMC管理	電力制御、ボード情報監視、ボードSNへのアクセスを含む インテリジェントなボード管理
アップグレード	PCIeオンラインファームウェアアップデート対応

inspur

F10A

- **技術仕様**

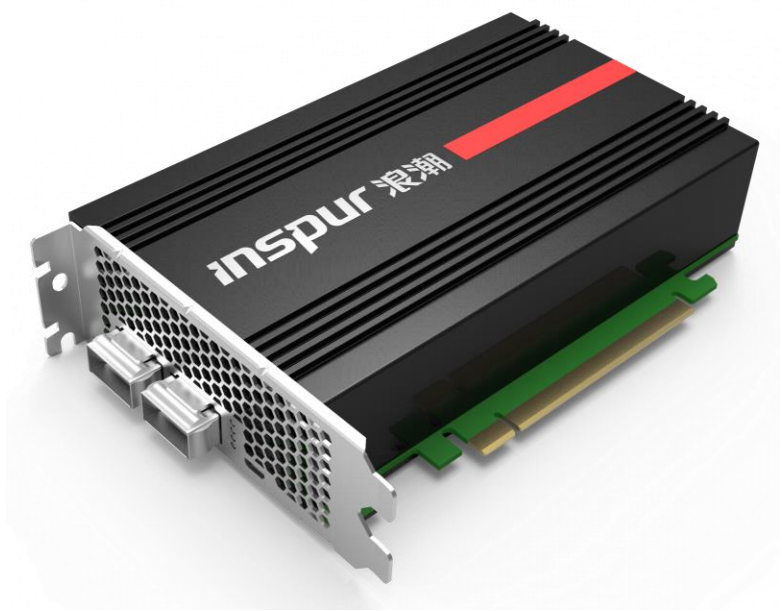
Model	F10A
サイズ	Half Length, Half- Height with low profile bracket Single width SKU1/Double width SKU2
システム インターフェイス	PCIe 3.0 x8 2*10GE interfaces
電　源	derived from on-board 12V PCIe slot
消費電力	45W(Peak)/35W(Average)
オンボード メモリ	Dual channel DDR4 SODIMM, 16GB @2133MT/s, 64bit
オンボード フラッシュ	2Gbit Flash with 32-bit width data bus
ピークパフォーマンス	1.366TFlops
プログラム言語	RTL, OpenCL, HLS
開発ツール	Intel OpenVINO SDK

2種類の放熱モデル

SKU1 - Active heat sink

SKU2 - Passive heat Sink

F37X: Ultimate Design for AI Computing



主要コンポーネント	
サイズ	Full height, half length (167mm x 111mm)
チップ	Xilinx Virtex UltraScale+ VU37P
メモリー	
HBM DRAM	8GB on-chip HBM2, bandwidth @ 460GB/s
DDR	support up to 24GB DDR4 @ 2,666Mb/s , 3 channels 72bits DDR4 SDRAM，
インターフェイス	
PCI Express	PCIE 3.0 x16
ネットワーク	2x 100GbE QSFP28
メンテナンスポート	Micro USB
電力 & 放熱	
電　源	12V PCIe slot @75W + additional 12V AUX supply @75W
消費電力	150W (Peak)，75W (Average)
冷却方法	Double-slot passive cooling
ボード管理	
ボード管理コントローラー (BMC)	電源の制御、ボード情報の監視（温度、消費電力、メモリ）、およびボードSNへのアクセスを含むインテリジェントなボード管理
アップデート方法	PCIeオンラインファームウェアアップデート対応

inspur

エンドトゥエンドの幅広いソリューションの提供

プログラミング言語

RTL, C, C++, OpenCL

Software IPs

webp

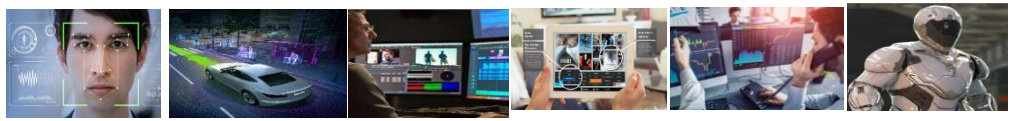


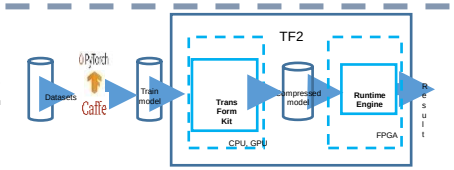
Image Recognition | Autonomous Driving | Video Processing | Video Recommendation | Financial Terminals | CAD/Content Creation

ML フレームワーク

TensorFlow | theano | torch | Caffe

開発ツール

TF2



OpenVINO™

XILINX
ML Suite for Inference

開 発 環 境

QUARTUS PRIME DESIGN SOFTWARE | HLS COMPILER | FPGA SDK FOR OPENCL | DSP BUILDER FOR INTEL FPGAS

VIVADO | SDAccel

ハードウェアプラットフォーム

Inspur FPGAs

NF5280M5

NF5468M5

inspur

Inspur FPGA成功ケース

- **Inspur FPGAビジネスの利点**

データセンター向けの総合的なFPGAシステムソリューション
成熟した信頼できる品質管理
アフターサービスとメンテナンス
ソフトウェア生産性

NF5280M5

F10A

Alibaba Cloud Ali FPGA Cloud	科大讯飞 iFLYTEK IFLYTEK Speech Recognition	Falcon COMPUTING Life Science acceleration
Tencent Cloud Tencent FPGA Cloud	Horizon Robotics Video decoding in Horizon Robotics	华大基因 BGI Life Science acceleration
金山云 WWW.KSYUN.COM KingSoft FPGA Cloud	網易 Transcoding in NetEase	華中科技大学 HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY Brainsmatics Research

inspur

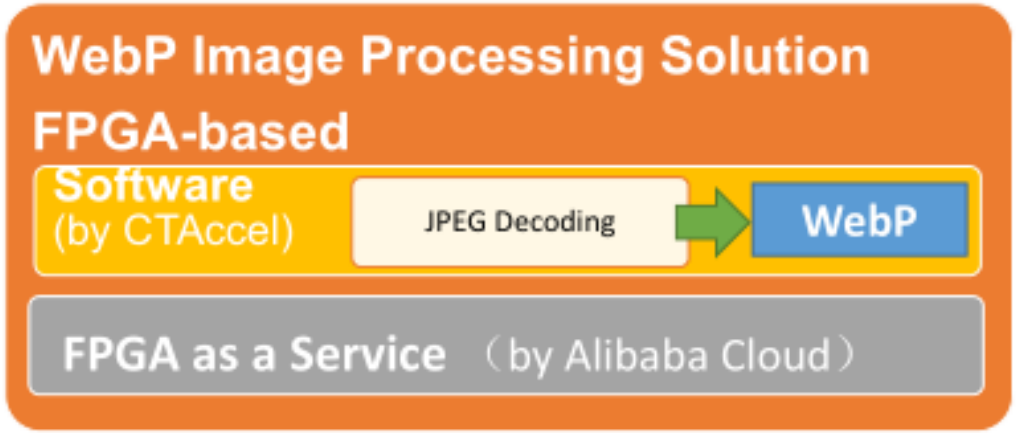
FPGA-based WebP image processing solution

Critical Challenge

Affected by the slow decoding speed of the average format of JPG and PNG, few companies in China are well able to use WebP, a image compression format meant to reduce the size of JPEG images.

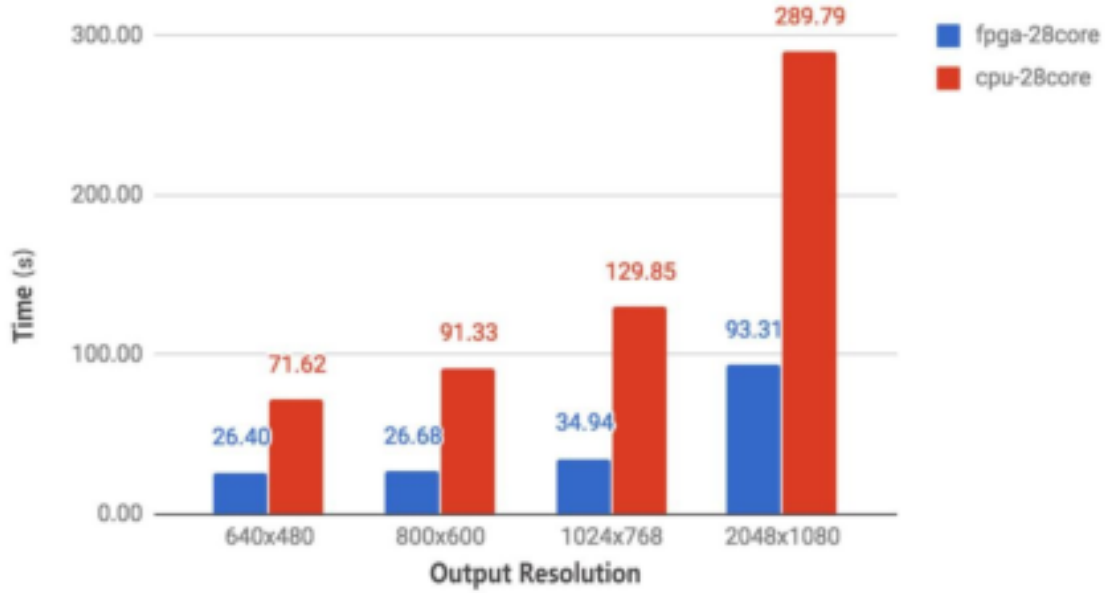
FPGA Solution

This solution further unleash the capability of FPGA in data parallel processing to meet the increasing computing demand, making image processing more efficient. It enables users to utilize JPGE decoding system and WebP coding system to convert JPEG format into WebP in batches.



User Benefit

This FPGA-based image process solution achieves several times of performance that of traditional image processing based on CPU.



Horizon Robotics

FPGA-based intelligent video analytics(IVA) solution

Horizon Robotics

Critical Challenge

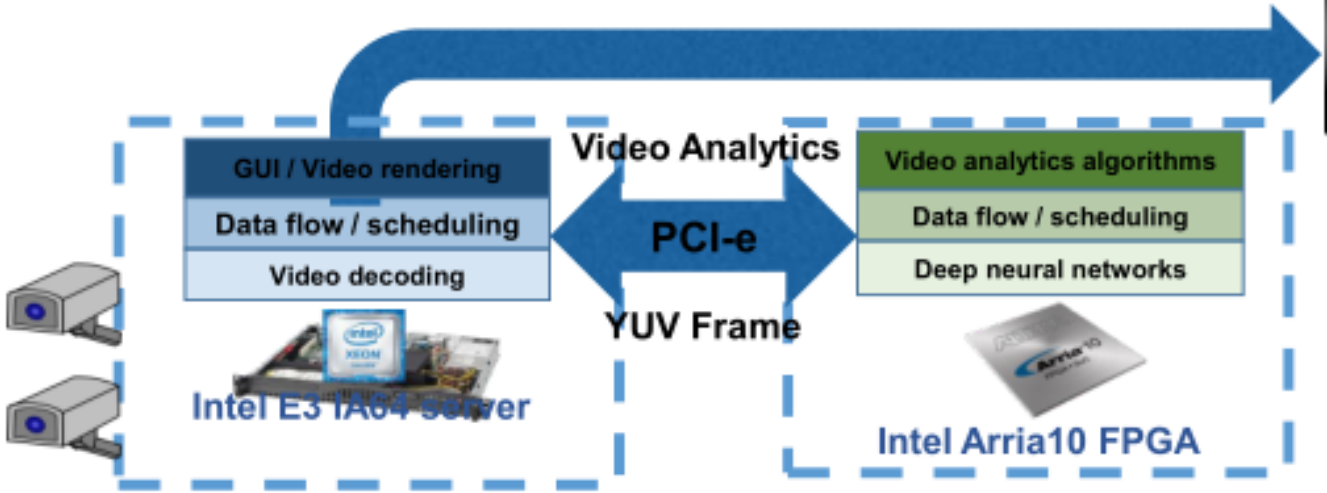
Major pain point is how to efficiently translate unstructured data (video) into insight for security and public safety.

User Benefit

The solution helps security and public safety organizations develop comprehensive security, intelligence, and investigative capabilities using video..

FPGA Solution

FPGA-based intelligent video analytics(IVA) solution with high density, high Performance and high reliability.
Performance Indicator: Video 1080P@25fps*16 channels



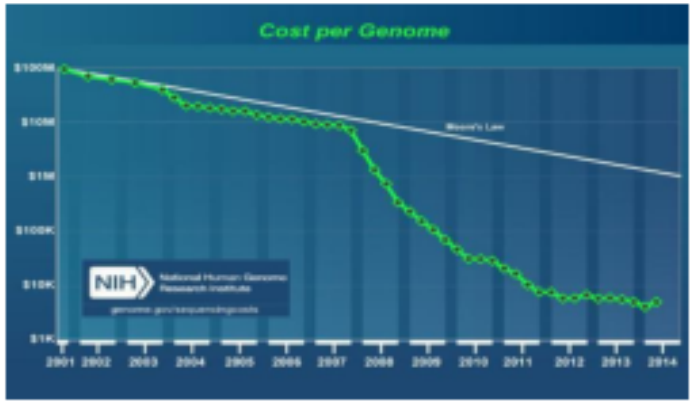
inspur

Falcon Computing

F10A acceleration for life sciences

Critical Challenge

- **Genome Sequencing cost is reducing faster than Moore's Law, but computing cost is not**



- **On-Premises and Cloud platforms**
  - On-premise: fixed space/energy budget, performance bottleneck
  - Cloud: long upload time, security, high cost for compute and storage

FPGA Solution

Differentiation, Approach for GATK Secondary Analysis

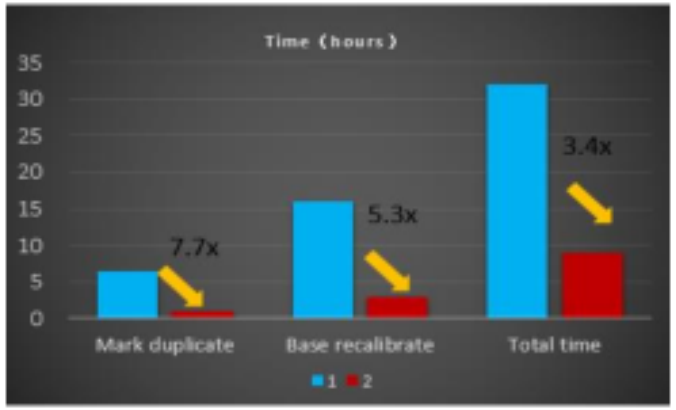
Use of Compiler and acceleration algorithms on standard, best-practices GATK pipeline

- Allows aggressive acceleration, rapid turn-around times, and a scalable approach
- Able to quickly support new code bases/releases or migrate across h/w platforms
- Enables use of FPGA accelerators without expertise in FPGA design or OpenCL languages
- Able to eventually scale to hybrid, heterogeneous clouds (CPU, FPGA and other)
- Continues to improve performance and support emerging h/w platforms

Falcon Genomics solution showing outstanding acceleration on one of its partner-Intel's Arria10 FPGA platform

User Benefit

This solution, supported by FPGA, provides considerable speed-up for life sciences applications.



Inspur GeneEngineM2 with Falcon Computing Genomics Solutions shows 3.4x speedup compared to GeneEngineM1

Picture transcoding solution -- WebP

In 2010, Google release that will support gif in 2014, it can compress the picture more and also can guarantee the image quality.

PNG： reduce more than **26%**　　JPG： reduce **25-34%**　　GIF： reduce **64%**

Google —— for all product line

- Page transmission reduce **33%**
- Page loading time reduce **10%**
- Save **TB level** storage space everyday

Tencent —— News client

- Bandwidth's peak of flow reduce **9GB**
- Delay of picture & delay of data's download reduce **100ms**

inspur