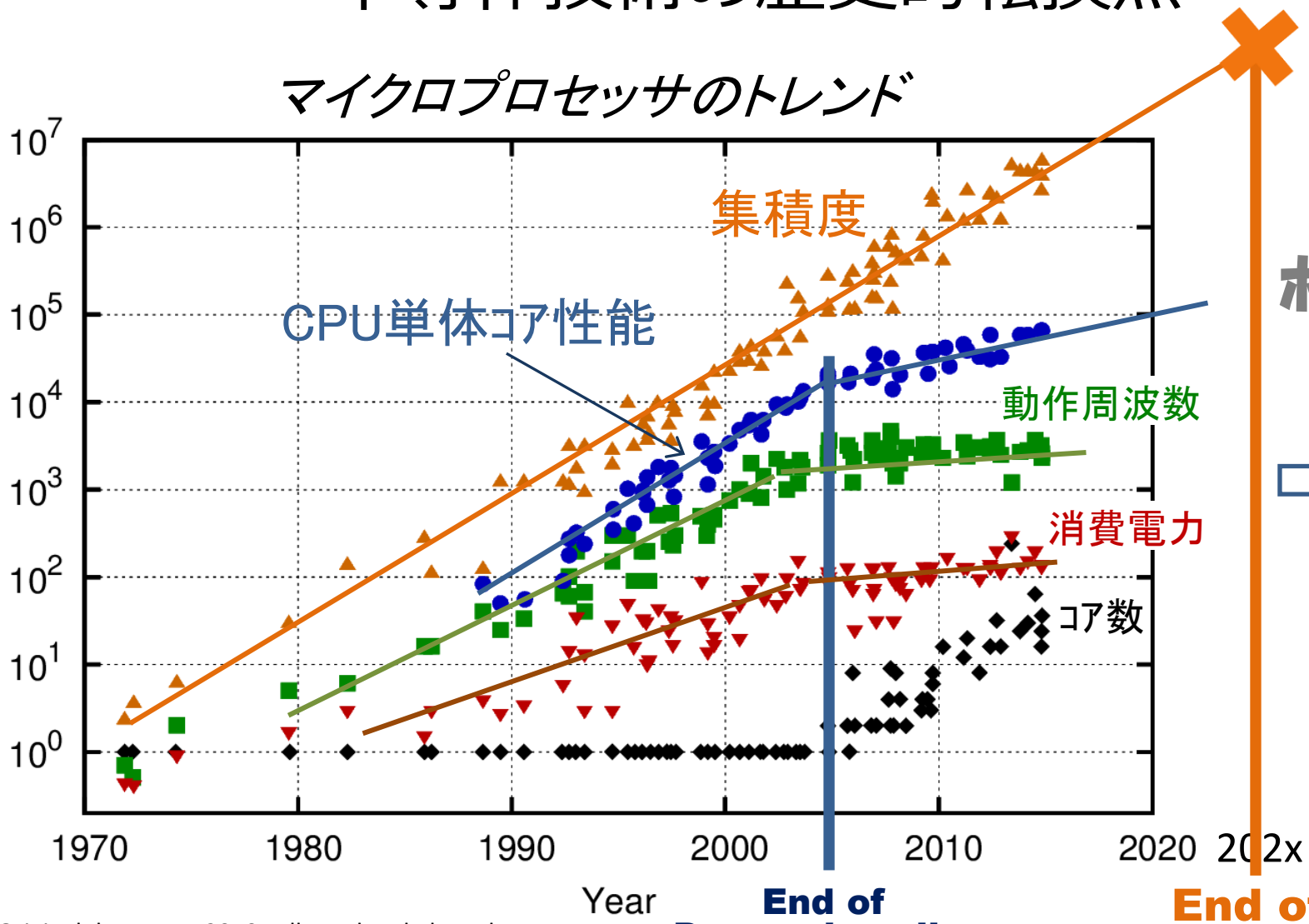
ポストムーアの半導体技術とAIチップ設計拠点の活動

産業技術総合研究所
内山邦男

１．ポストムーアの半導体技術

２．AIチップ設計拠点の活動

半導体技術の歴史的転換点



ポストムーア時代
（202x年～）

End of Dennard scaling

- 寸法：1/k
- スイッチング速度：k
- 消費電力：$1/k^2$

End of Moore's Law ?

集積度：18ヵ月で2倍

Original data up to 2010 collected and plotted by M.Horowitz, F.Labonte, O.Shacham, K.Olukotun, L.Hammond, and C.Batten
New plot and data collected for 2010-2015 by K.Rupp

ムーアの法則の終焉後に
「指数関数的な性能向上」を担うものは何か？

2012　IEEE Rebooting Computing Initiativeがスタート
**“to rethink the computer, "from soup to nuts,"
including all aspects from device to user interface.”**

2015　次世代スパコンに向けた大統領令発令
戦略目標の１つに**“ﾎﾟｽﾄﾑｰｱに向けた研究開発の強化”**を設定

IARPA(情報機関系の研究支援組織)が量子計算、JJ計算機、
ﾆｭｰﾛ計算のﾌﾟﾛｼﾞｪｸﾄ推進強化

2016　Googleが**TPU**発表 (TPU2.0 in 2017, TPU3.0 in 2018)

2017　AppleがiPhone用A11チップに**Neural Engine**搭載

DARPAが**Electronics Resurgence Initiative**ﾌﾟﾛｼﾞｪｸﾄ開始

2018　Patterson教授のTuring Awardレクチャー@ISCA
“A New Golden Age for Computer Architecture”

International Symposium on Computer Architecture, 2017

In-Datacenter Performance Analysis of a Tensor Processing Unit™

Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon

Many architects believe that major improvements in cost-energy-performance must now come from **domain-specific hardware**.

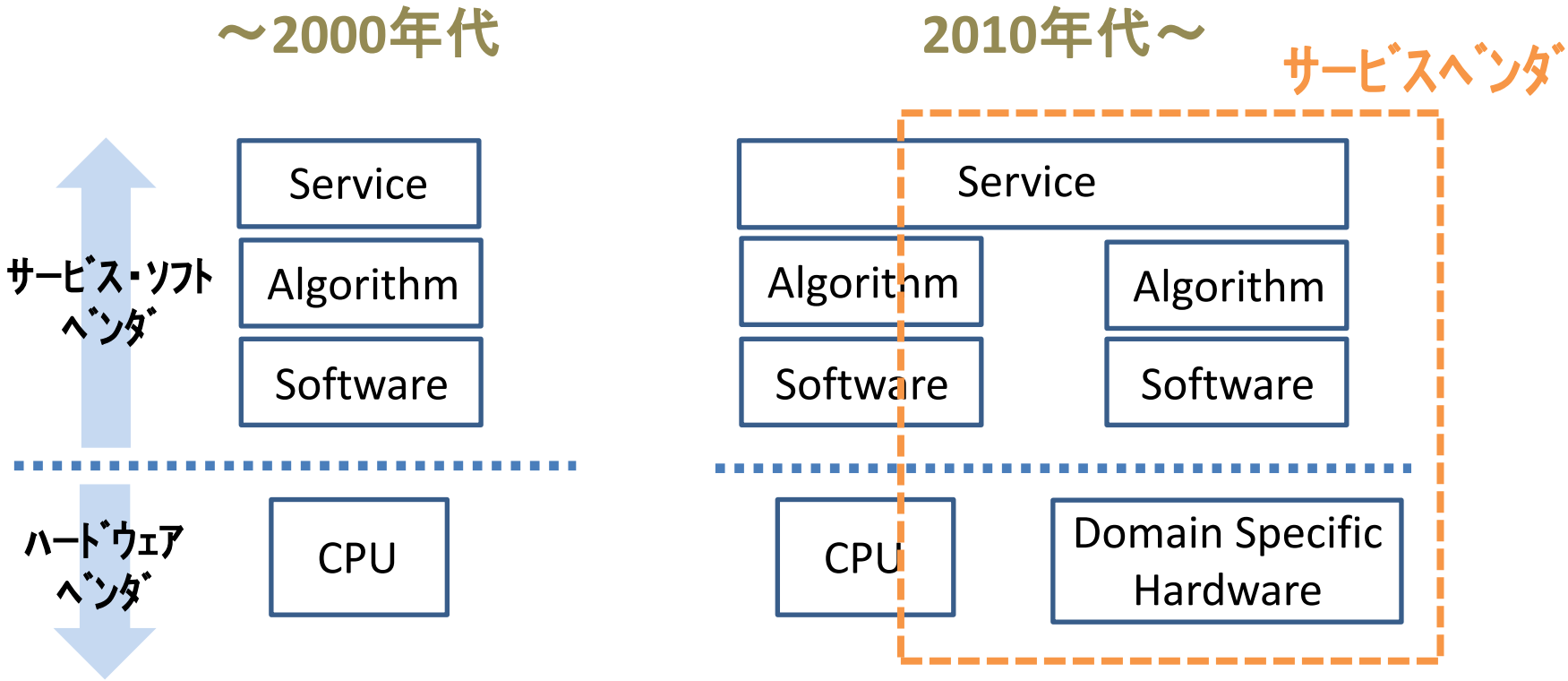
To appear at the 44th International Symposium on Computer Architecture (ISCA), Toronto, Canada, June 26, 2017.

Abstract

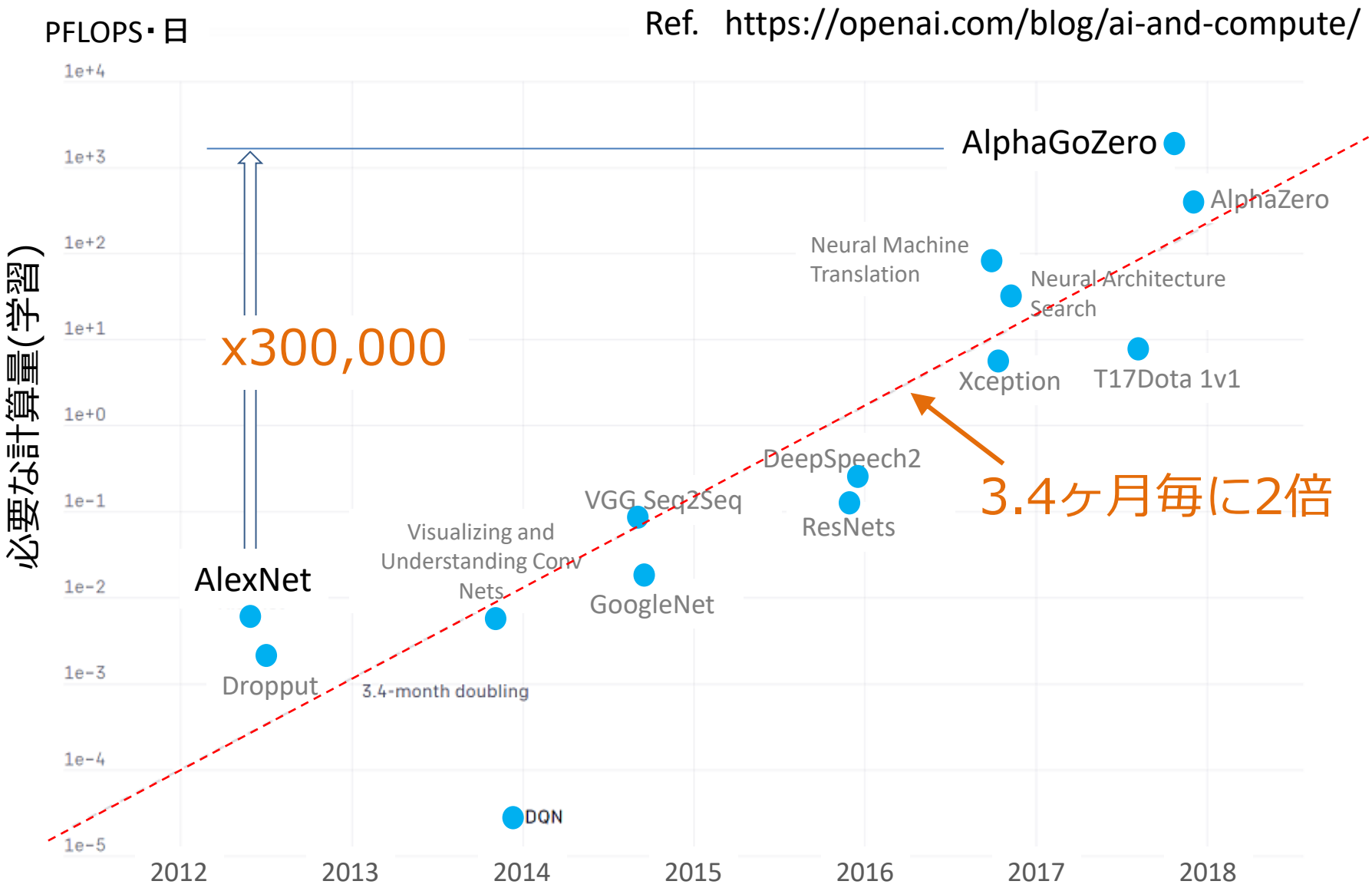
Many architects believe that major improvements in cost-energy-performance must now come from domain-specific hardware. This paper evaluates a custom ASIC—called a *Tensor Processing Unit (TPU)*— deployed in datacenters since 2015 that accelerates the inference phase of neural networks (NN). The heart of the TPU is a 65,536 8-bit MAC

D.Patterson教授のTuring Award受賞講演@ISCA2018

"A New Golden Age for Computer Architecture: Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development"

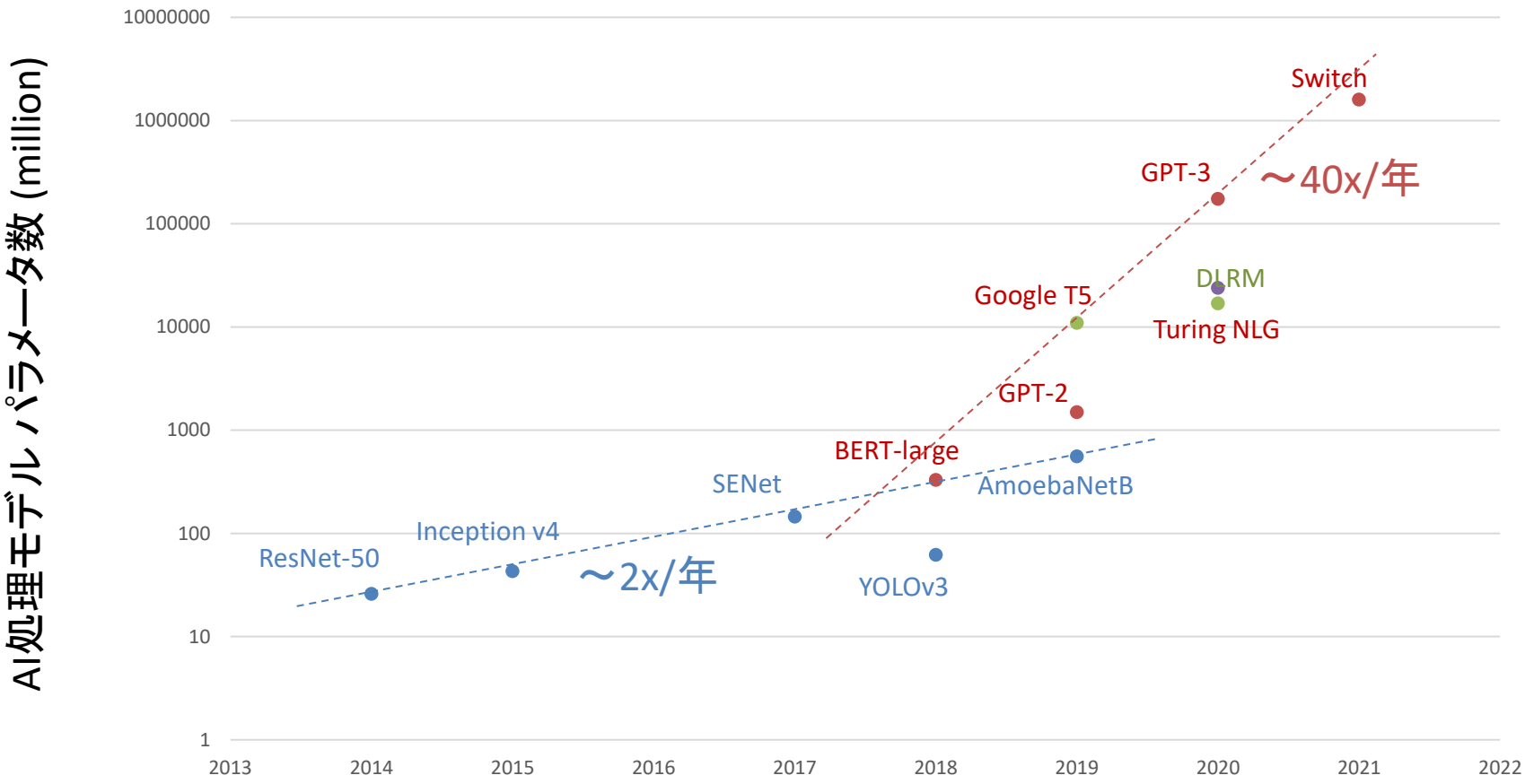


AI学習($y_i = \sum_j w_{ij} x_j$)の計算量



AI処理モデル($y_i = \sum_j w_{ij} x_j$)のパラメータ数

Ref. Keynote by Linley Gwennap,
Linley Fall Processor Conference 2021

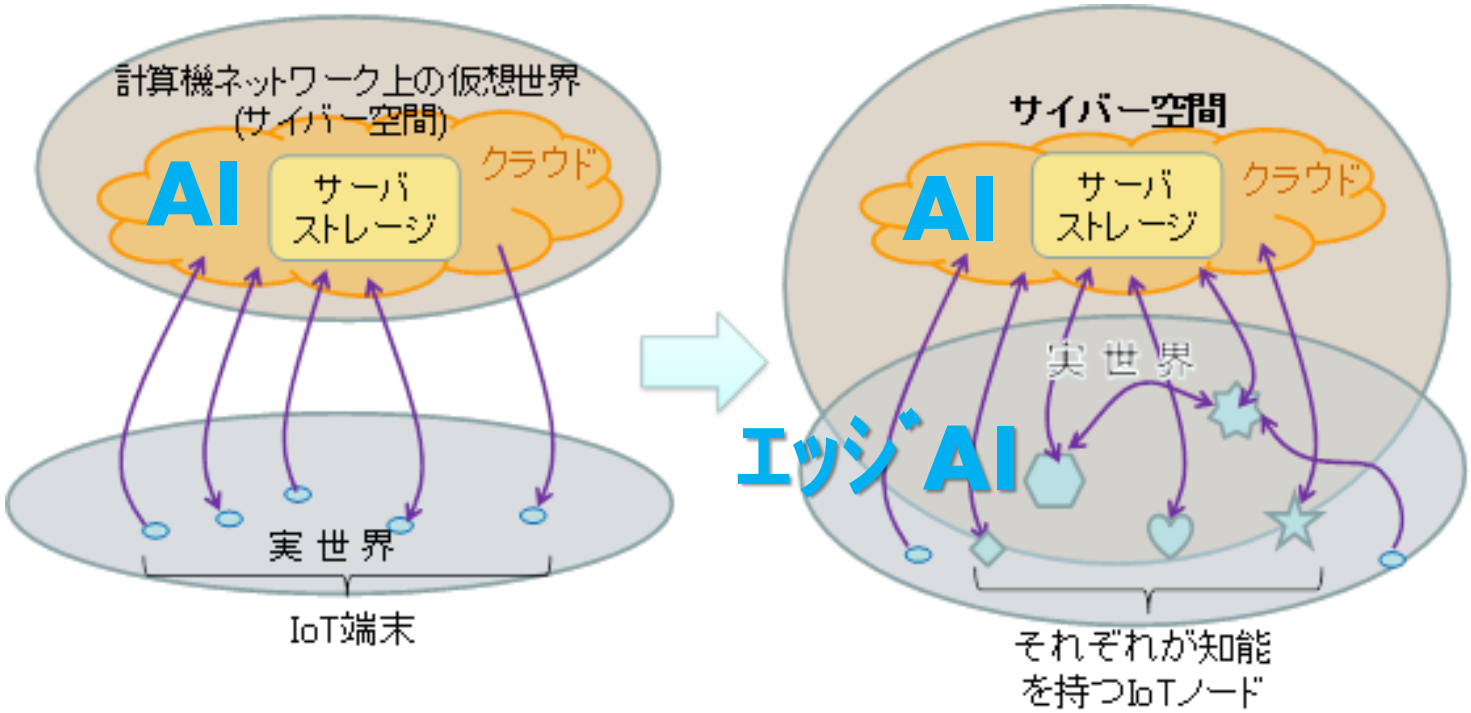


サービス事業者によるAIチップ開発

- **Google**
  AIチップ(TPU)を自社開発 (2015, 2017, 2018)
  自動翻訳、WEB検索などのクラウドサービスに活用
- **Apple**
  iPhone用SoCを自社開発 (A4～A11, 2010～)
  A11(2017)で、AIアクセラレータを搭載
- **Facebook**
  自然言語処理（NLP）に特化したAIチップ開発中(2019/ISSCC)
- **Amazon**
  推論専用AIチップ(AWS Inferentia)を開発(2018)
  450名のチップ開発チーム
- **Baidu**
  クラウド向けAIチップ(Kunlun)を開発 (2017, 2020/Hot Chips)
  14nm, HBMx2, 150W, 256TOPS, PCIeGen4x8
- **Alibaba**
  推論専用AIチップ(Hanguang 800)を開発中 (2020/Hot Chips)
  12nm, 825TOPS, 192MB on-chip SRAM, 700MHz
- **Preferred Networks**
  学習用AIチップ(MN-Core)を開発（2018, 2020/ISC)
  524TFLOPS, 500W/4die

クラウドから**エッジ**へ

「2030年に向けた次世代計算機技術開発戦略」 日本工学アカデミー報告書



「AI Moves From the Cloud to the Edge」

基調講演 by Linley Gwennap, Linley Fall Processor Conference, October 20, 2021

エッジAIチップの状況

Ref. Linley Fall Processor Conference, 2021

- ハイエンド（自動運転、etc.）
  - **Nvidia Orin X**：72 TOPS at 25W～254 TOPS at 100W
  - **Xilinx Versal AI Edge FPGA**：202 TOPS at 75W
  - **Qualcomm Cloud AI 100**：100 TOPS at 15W～400 TOPS at 75W
- ミドル/ローエンド（監視カメラ、etc.）
  - **Ambarella**：12 TOPS accelerator at 5W
  - **Quadric**：4 TOPS at 2W
  - **ArchiTek**：4 TOPS at 1.5W
  - **RealTek**：0.7 TOPS at 0.8W
- 超低電力（IoTセンサ、etc.）
  - **Syntiant**：500uW
  - **Innatera**：100uW
  - **Ambient, BrainChip**：サブmW

AIチップ関連ベンチャ/SME

Horizon Robotics(中国)
Cambricon Technologies(中国)
Cerebras Systems(US)
Graphcore(UK)
Wave Computing(US)
Hailo(イスラエル)
Mythic(US)
Kneron(US)
Flex Logix Technologies(US)
DeePhi(中国)
NSITEXE(日本)
Syntiant(US)
Cornami(US)
TRIPLE-1(日本)
SiMa.ai(US)
GrAI Matter Labs(仏)
brainchip*(US)
Eta Compute(US)

Movellus, Inc.(US)
Greenwaves Technologies(仏)
innatera Nanosystems(オランダ)
Groq(US)
edgeCortix(日本)
LeapMind(日本)
Think Silicon(ギリシャ)
Perceive(US)
Alphawave(カナダ)
Expedera(US)
Gyrfalcon Technology(US)
ArchiTeK(日本)
Digital Media Professionals(日本)
Axell(日本)

各国の半導体関連強化施策

第4回 半導体・デジタル産業戦略検討会議
「半導体戦略の進捗と今後」資料より

国・地域	産業支援策の主な動向
米国	• **最大3000億円/件の補助金や「多国間半導体セキュリティ基金」設置等**を含む国防授権法（NDAA2021）の可決。 • バイデン大統領はCHIPS法案に賛意を表明。上院においては**5.7兆円**の半導体関連投資を含む**「米国イノベーション・競争法案」**が通過。
中国	• 「国家集積回路産業投資基金」を設置('14, '19年)、**半導体関連技術へ、計５兆円を超える大規模投資**。 • これに加えて、地方政府で**計5兆円を超える半導体産業向けの基金**が存在（**合計10兆円超**）
欧州	• 2030年に向けたデジタル戦略を発表。**デジタル移行(ロジック半導体、HPC・量子コンピュータ、量子通信インフラ等)に1345億€(約17.5兆円)投資**等 • 製造を含む欧州の最先端チップ・エコシステムの構築を目指し、供給の安全を確保し、欧州の画期的技術のための新たな市場を発展させる**「新・欧州半導体法案」**の制定を宣言（2021.9）
台湾	• **台湾への投資回帰を促す補助金等の優遇策**を始動。ハイテク分野を中心に**累計で2.7兆円の投資申請**を受理。（2019.1） • **半導体分野に、2021年までに計300億円の補助金**を投入する計画発表。（2020.7）
韓国	• **AI半導体技術開発への投資**に**1,000億円**を計上。（2019.12） • **半導体を含む素材・部品・装置産業の技術開発**に2022年までに**5,000億円以上を集中投資**する計画を発表。(2020.7) • 総合半導体大国実現のための「**K-半導体戦略**」を策定（2021.5）

AIﾁｯﾌﾟ/次世代ｺﾝﾋﾟｭｰﾃｨﾝｸﾞ/半導体関連の施策（2018年～）

- **経産省関連**
  - AI チップ開発加速のための検証環境整備事業
  - AIチップ・次世代コンピューティングの技術開発事業
  - AIチップ開発加速のためのイノベーション推進事業
  - ポスト5G情報通信システム基盤強化研究開発事業
  - グリーンイノベーション基金事業

- **文科省関連**
  - Society5.0を支える革新的コンピューティング技術
  - 革新的コンピューティング技術の開拓
  - 次世代X-nics半導体創生拠点形成事業

２．AIチップ設計拠点の活動

NEDO事業

「AIチップ開発加速のためのイノベーション推進事業

研究開発項目②：AIチップ開発を加速する共通基盤技術の開発」

https://www.nedo.go.jp/activities/ZZJP_100142.html

拠点構築の目的

- ✓ 我が国では、ベンチャー企業等を中心に、AIチップを基にした新たなビジネスを創出させる種が多数存在。
- ✓ 一方、AIチップ設計には、高額な**EDAツール**や**IP**、**検証装置（エミュレータ等）**が必要であり、これらがビジネス化に向けた高いハードルとなっている。
- ✓ AIチップ設計に必要な設計・検証環境を整備し、イノベーション実現のためのAIチップ開発を加速する。

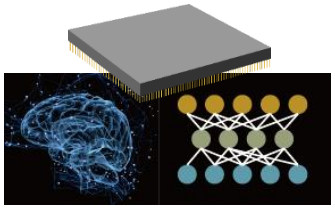
革新的AIチップのアイディア

国内中小企業
ベンチャー企業

高額なEDAツール、IP、検証装置が必要

高いハードル

AIチップ プロトタイプ試作



学習、推論、認識を低電力かつ高速に

AIチップ設計拠点

超スマート社会（Society5.0）の実現

- **次世代モビリティ**
  自動運転，無人配送，…
- **次世代ヘルスケア**
  AI診断，自動ﾓﾆﾀﾘﾝｸﾞ，…
- **次世代サプライチェーン**
  ｽﾏｰﾄ保安，無人工場，…
- **農林水産業スマート化**
  無人農業車両，水中ﾛﾎﾞｯﾄ，…
- **FinTech**

：

拠点の体制・運営

赤字：AIチップ設計拠点 運営組織

ベンチャーキャピタル
Pluga Capital

ベンチャー・中小企業 等

EDAツール・IP・検証環境提供

サテライト拠点

ふくおかIST

AIチップ設計拠点
@東大本郷地区浅野キャンパス

産総研

東京大学

産総研

人工知能研究センター
ABCI

大学・公的機関
コンソーシアム

・北海道大学
・東北大学
・福岡大学

EDAツール ベンダー

IP ベンダー

ファウンドリ

LSIデザイン ハウス
凸版印刷, …

ソフトウェア ハウス

：

✓ AIチップ設計に必要な設計・検証環境
（EDAツール，IP，エミュレータ等）の整備，提供
✓ AIチップ開発に資する設計技術、検証手法の開発
✓ AIチップ技術に関する人材育成

EDAツール

Cadence, Synopsys, Siemensのツールを整備

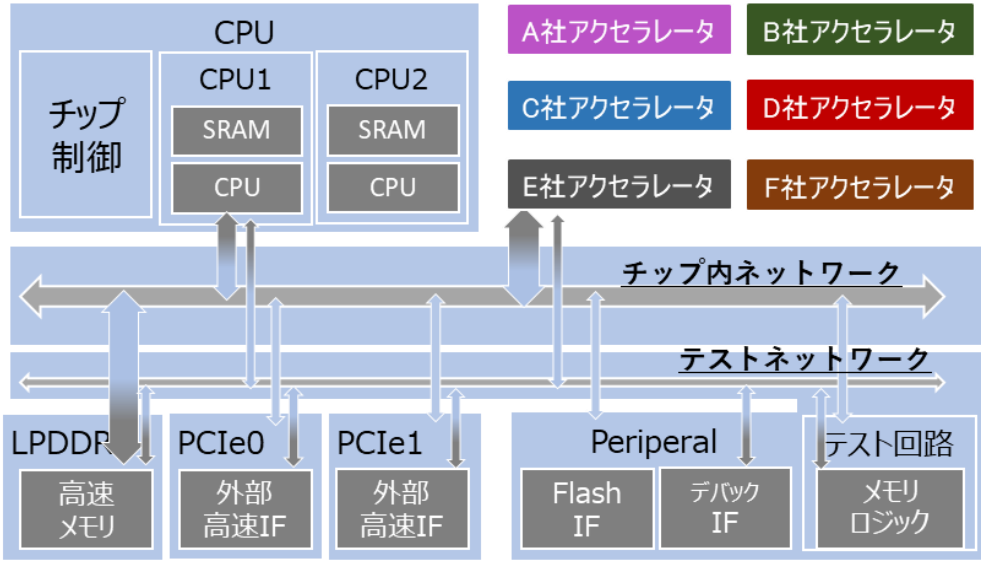
アーキテクチャ検証、高位合成、論理設計/検証、回路設計、物理設計/検証、
論理エミュレータ、FPGAプロトタイピング、ボード設計、etc.

IP

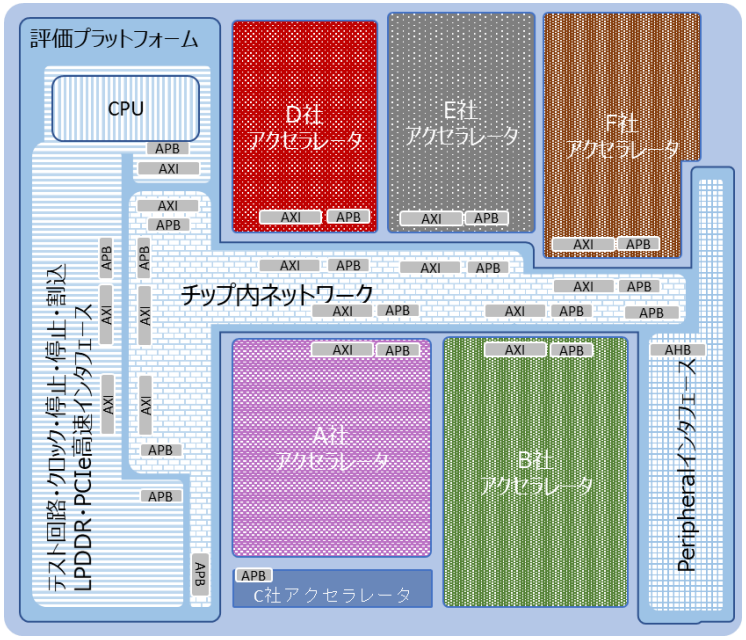
Synopsysの40nm, 28nm, 12nm 標準IP群
(CPU, DSP, DMAC, DDR, PCIe, USB, MIPI, I2C, etc.)
を整備
(物理系はTSMC向け)

エッジAI向け評価チップ (AI-One)

- 拠点導入IP(28nm)を活用して、拠点が評価プラットフォームを準備
- 乗合チップ参加企業は各社のAIアクセラレータをプラットフォームに接続
- 拠点がまとめてチップ実装を行いファブに試作依頼、各社にチップ(+ボード)を配布
- 各社は試作チップ（ボード)を用いて、実証実験を実施中



チップ内部構成



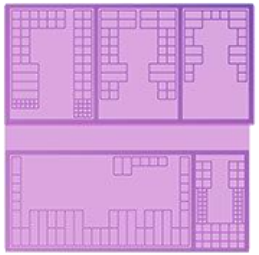
チップ実装イメージ

2022年3月プレスリリース

複数のAIアクセラレータを搭載した実証チップ「AI-One」の動作を確認
–従来比45%以下の短期間で低コストのAIチップ設計・評価が可能に–

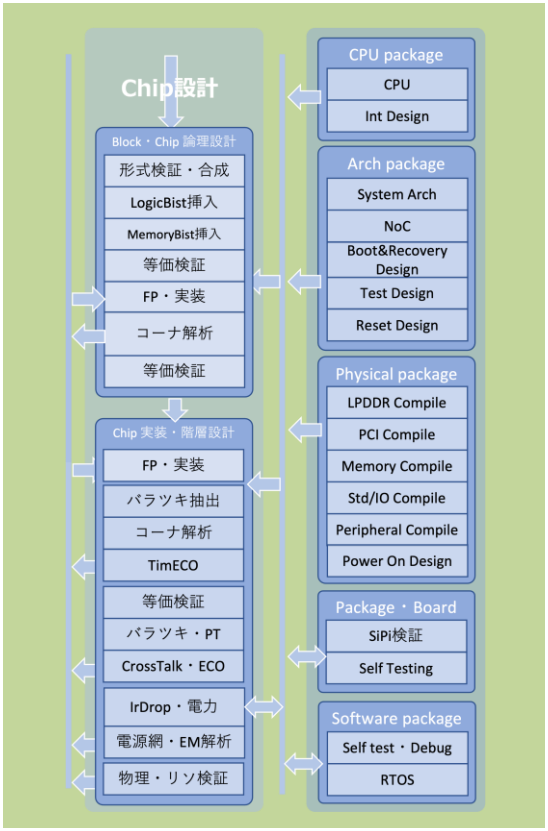
AI-One 設計/評価プラットフォーム

AIアクセラレータ

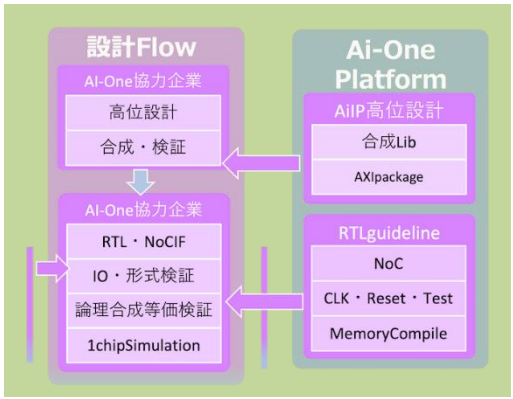


標準SoC回路

SoC設計手法

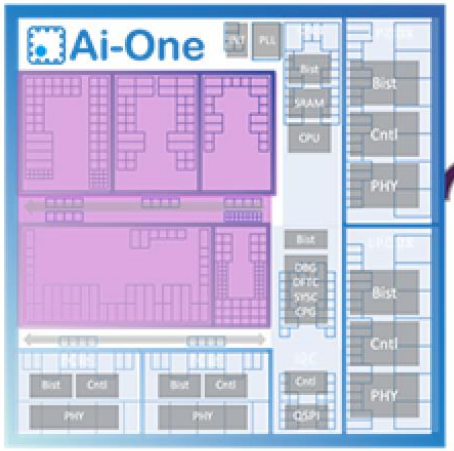


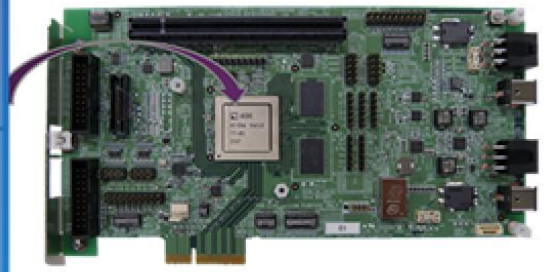
アクセラレータ設計手法



SoC化

SoC型AIチップ



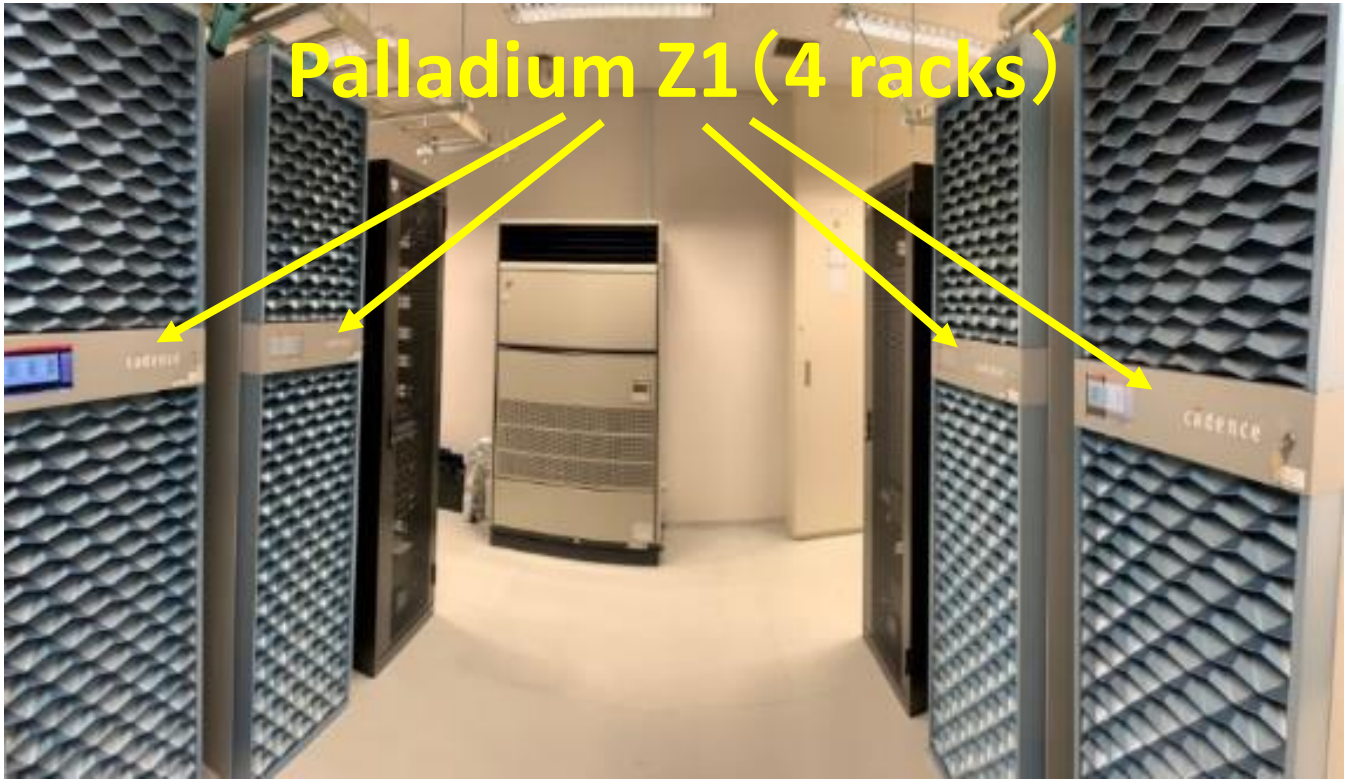


評価ボード+基盤ソフト

論理エミュレータ

Palladium Z1

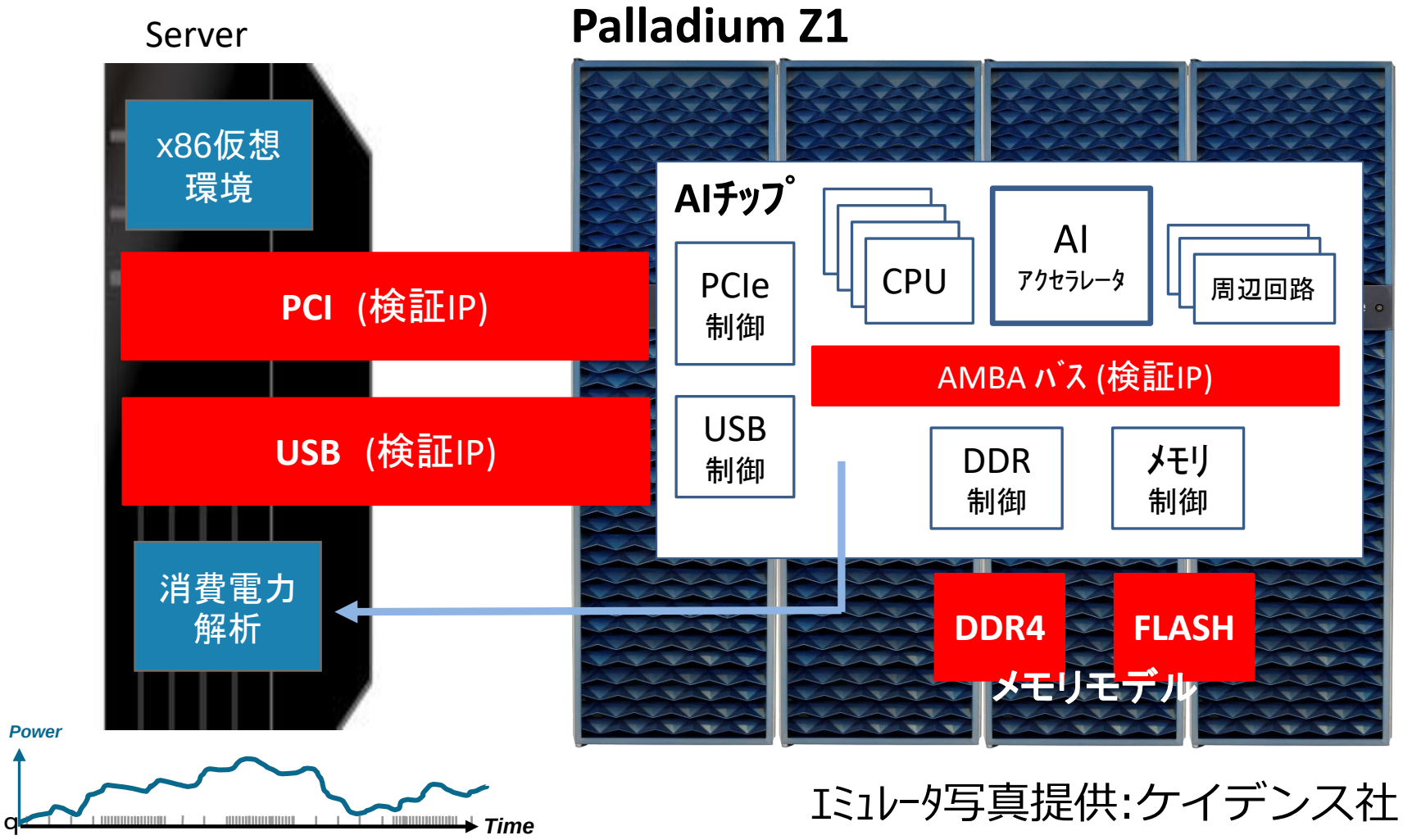
・容量：23億ｹﾞｰﾄ，4.6Tﾊﾞｲﾄ（ﾕｰｻﾞﾒﾓﾘ），4.6Tﾊﾞｲﾄ（ﾃﾞﾊﾞｯｸﾞﾒﾓﾘ）
・ｴﾐｭﾚｰｼｮﾝ速度：最大4MHz，ｺﾝﾊﾟｲﾙ速度：140Mｹﾞｰﾄ/時



AIチップ設計拠点 サーバ室

論理エミュレータの活用

論理シミュレーション(ソフト)に対して4,000倍～15,000倍の高速化



エミュレータ写真提供:ケイデンス社

RISC-V on 論理エミュレータ(Palladium Z1)

RISC-V Days Tokyo 2021 Spring, 2021/4/23, 講演

荒川 文男

東京大学 大学院 工学系研究科 システムデザイン研究センター

『Emulator上でのRISC-Vプロセッサによる Linuxブート』

- Cadence社のEmulator上でRISC-VコアによるLinuxのブートに成功
- RISC-Vコアは東工大一色研の高位システム設計検証環境で開発したコア
- Emulation環境構築では、UARTトランザクタを活用して、Linux端末ウインドウおよびEmulation状況のモニタリングを実現
- 約132M cycleの実行時間は最速で2分20秒程度
- 速度はCPU時間で約1.2M cycles/s、実行時間で0.9～1.1M cycles/s

EmulationとSimulationの比較

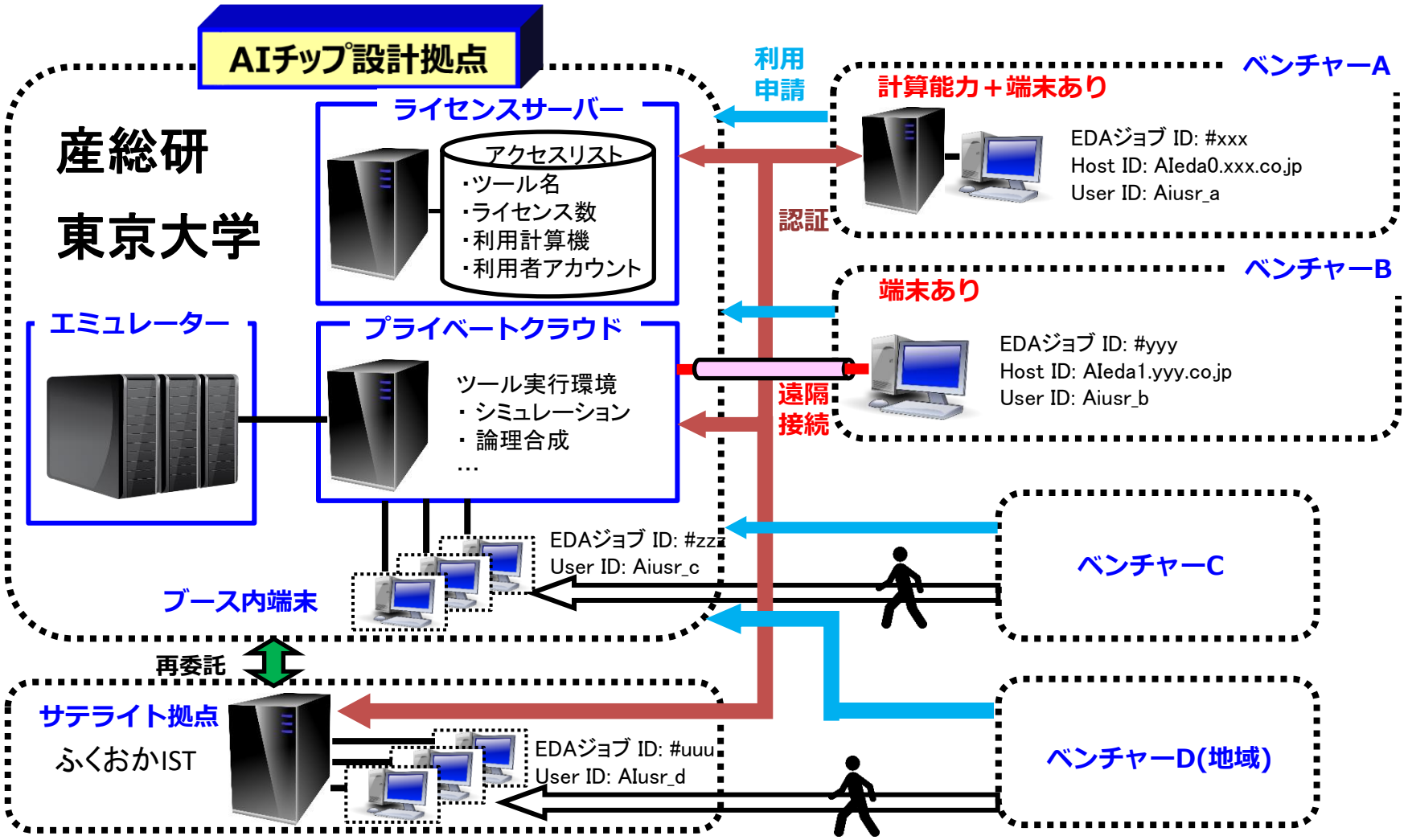
	parameters			boot cycles		Linux Boot			1M cycles		
	ext. mem.	log check	**UART**			RTL	Gate	G/R	RTL	Gate	G/R
1	log	Yes	1/2	116M	Emu.	13'38"	15'06"	1.11	0'46"	2'01"	2.63
					Sim.	*262:46'20"*	*561:47'40"*	*2.14*	2:15'55"	4:50'35"	2.14
					S/E	*1,156*	*2,232*		177	144	
7	mem.	No	**1/234**	132M	Emu.	**2'20"**	**2'31"**	**1.08**	0'22"	0'53"	2.41
					Sim.	***145:09'48"***	***609:19'36"***	**4.20**	1:05'59"	4:36'58"	4.20
					S/E	***3,733***	***14,527***		180	314	

– *Italic numbers are estimated values with extrapolation of 1M-cycle executions.*

- 1M cyclesのEmulationでは初期化時間が無視できず、Linux Bootの1/100未満の実行に1/18～1/3の時間がかかる。したがって、**1M cyclesでの比較は不適切**。
- 一方、Simulationでは初期化時間が無視でき、**外挿によるLinux Boot時間推定**は正しいと期待できる。
- UART=1/2設定の影響は変化する。Emulationでは大、SimulationのRTLでは小、Gateではほぼ無し。変化の少ない**UART=1/234での比較**が妥当。
- 赤字部分がこれに相当し、**RTLで3,700倍、Gateで14,500倍高速化**。

拠点利用形態

- ✓ 企業毎の設計環境に応じた拠点利用形態を整備し、中小・ベンチャー等の企業群が使い易い拠点を目指す



AIチップ設計拠点フォーラム

- ✓ AIﾁｯﾌﾟ、次世代ｺﾝﾋﾟｭｰﾃｨﾝｸﾞ、LSI設計などに関する技術情報を共有し、議論する場を提供
- ✓ 月１回のペースで開催（2019/5～）

第31回 AIチップ設計拠点フォーラム（2022/1/28）

13:30-13:35 **AIチップ設計拠点フォーラムについて**
（産総研／内山邦男）

13:35-14:35 **HotChips2021にみる機械学習アクセラレータの動向**
（東工大／本村真人先生）

14:35-15:35 **半導体デジタル産業戦略と関連施策について**
（経済産業省 ／齋藤尚史氏）

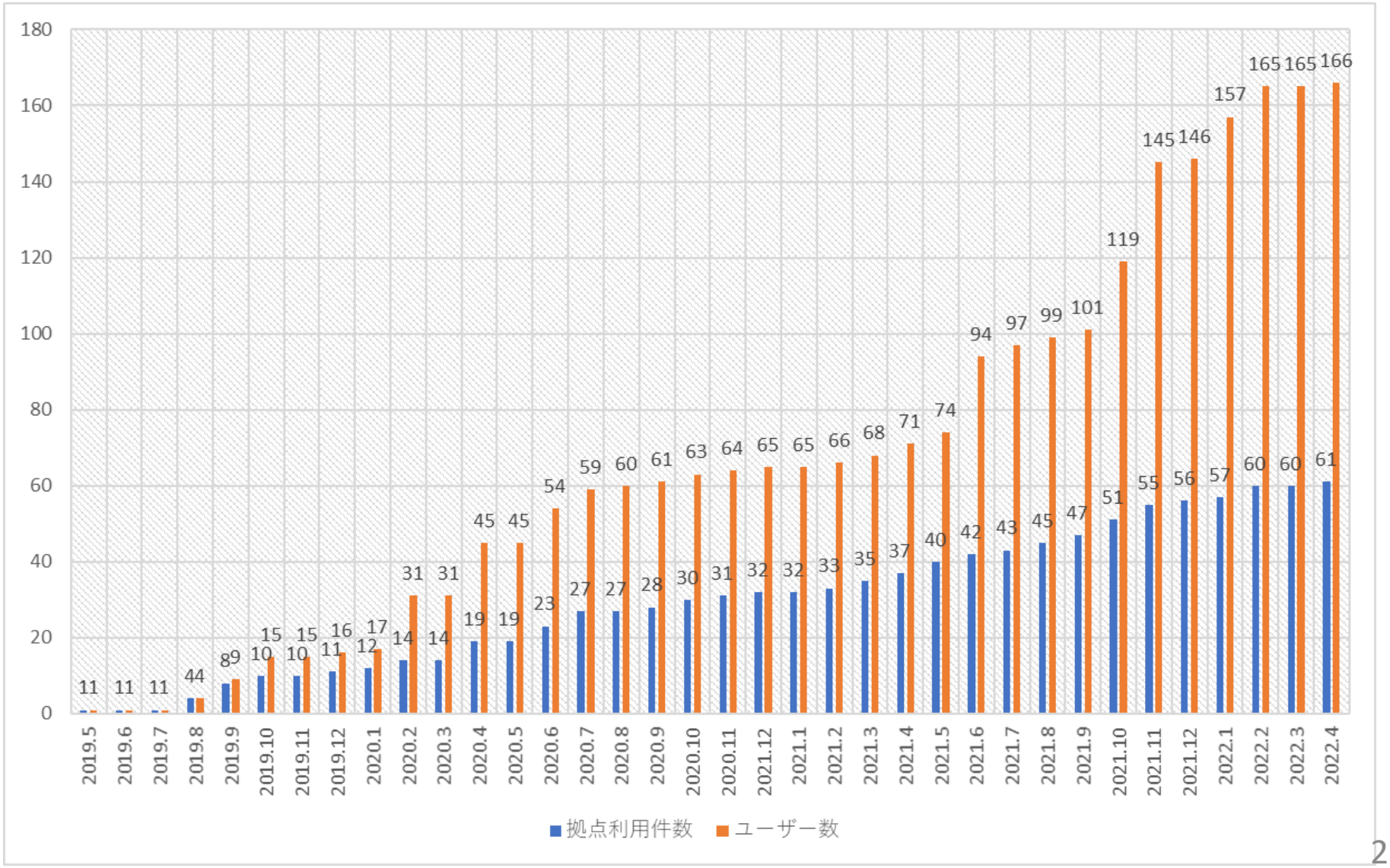
15:40-16:40 **次世代ロジック半導体に対応したオープンイノベーション拠点の整備**
（産総研／林喜宏氏）

拠点ＨＰ（https://www.ai-chip-design-center.org/）

Webinarと教材Download

拠点利用件数/ユーザ数の推移

利用件数 (2022/4月時点)：61件（企業:43件、大学:11件、国研:6件、その他:1件）



拠点ホームページ・拠点コンタクト先

https://www.ai-chip-design-center.org

AIチップ設計拠点事務局

- TEL：03-5841-8460
- 住所：〒113-0032
  東京都文京区弥生2-11-16武田先端知ビル203号室