AIチップ設計拠点[注)]

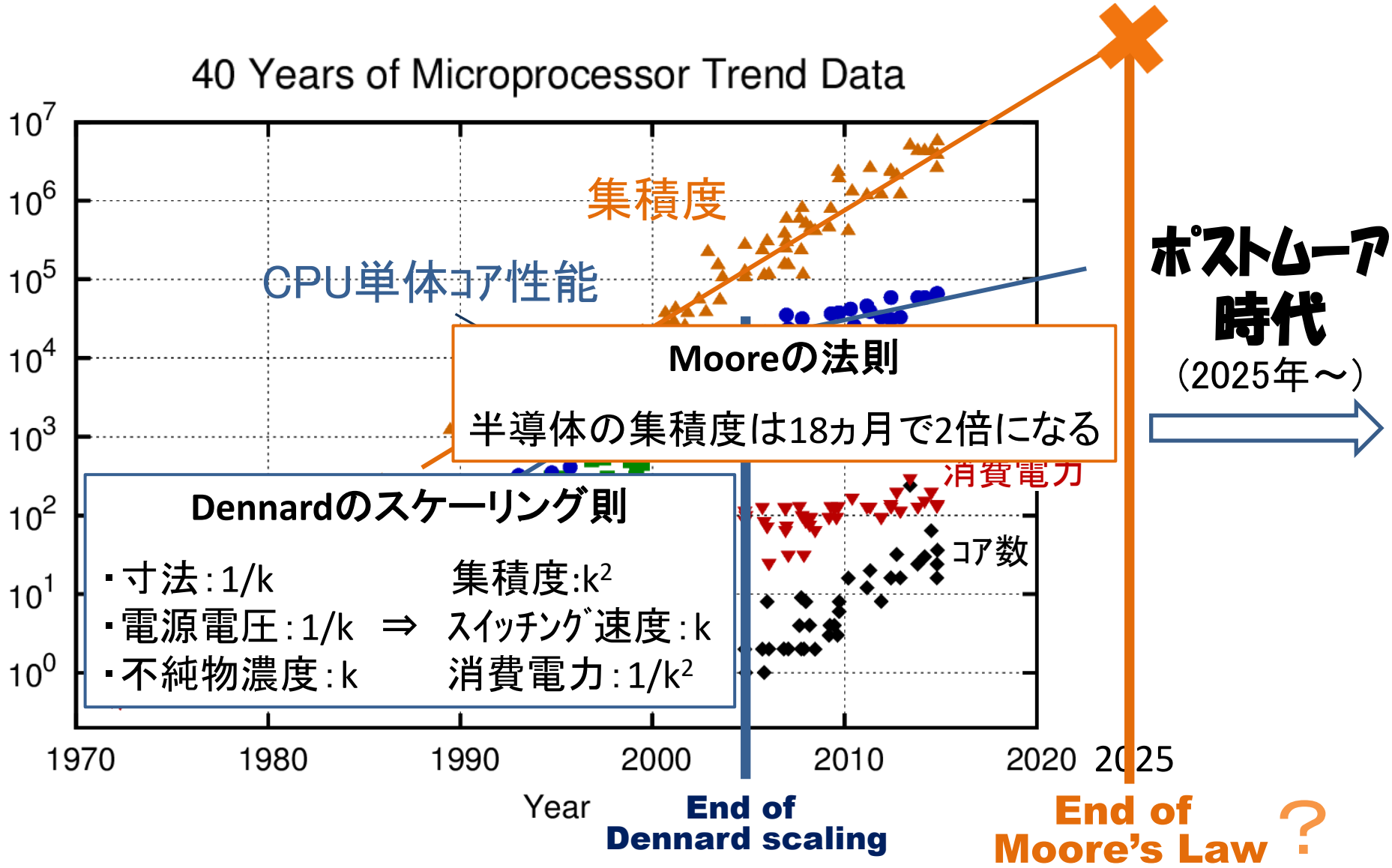
-AIチップに関連した世界の動きと拠点の活動-

産業技術総合研究所
内山邦男

注）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」

１．AIチップに関連した世界の動き

半導体技術の歴史的転換点



Original data up to the year 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”

IEEE International Conference on Rebooting Computing

- 2016より毎年１１月開催
- 2017年のセション構成
  - Neuromorphic Computing track
  - Probabilistic and Near-memory Computing track
  - Energy-efficient and Adiabatic Computing track
  - Novel Architectures and Near-memory Computing track
  - Quantum and Special Purpose Annealers track
  - Quantum Computing track
  - Optical Computing track
  - Algorithms and Applications track
  - Future EDA track
  - Beyond CMOS track
- 参考：
  - **Intel** Creates **Neuromorphic** Research Community to Advance **‘Loihi’** Test Chip, 1 Mar. 2018
  - **Intel** Starts R&D Effort in **Probabilistic Computing** for AI, 10 May 2018

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, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon

Google, Inc., Mountain View, CA USA

Email: {jouppi, cliffy, nishantpatil, davidpatterson}@google.com

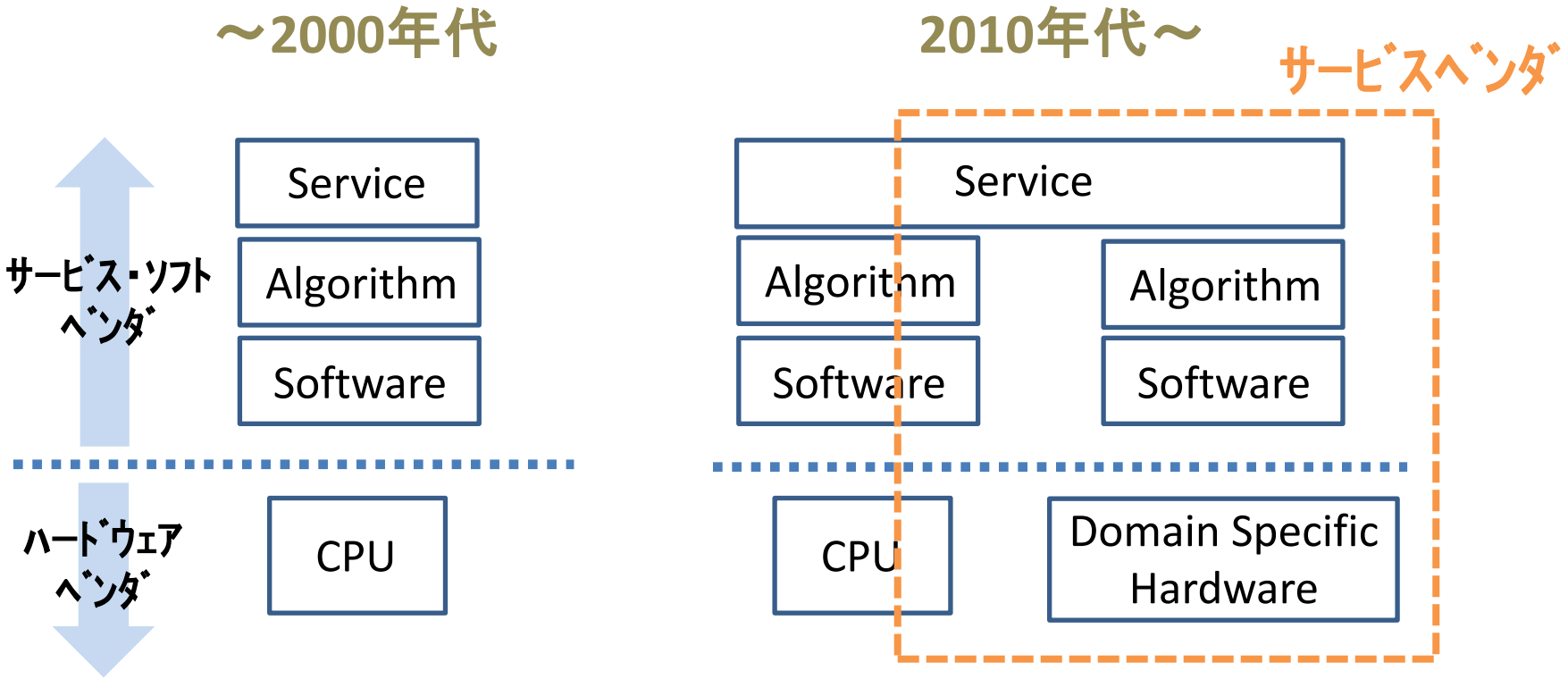
To appear at the 44th International Symposium on Computer Architecture (ISCA), Toronto, Canada, June 26, 2017.

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"



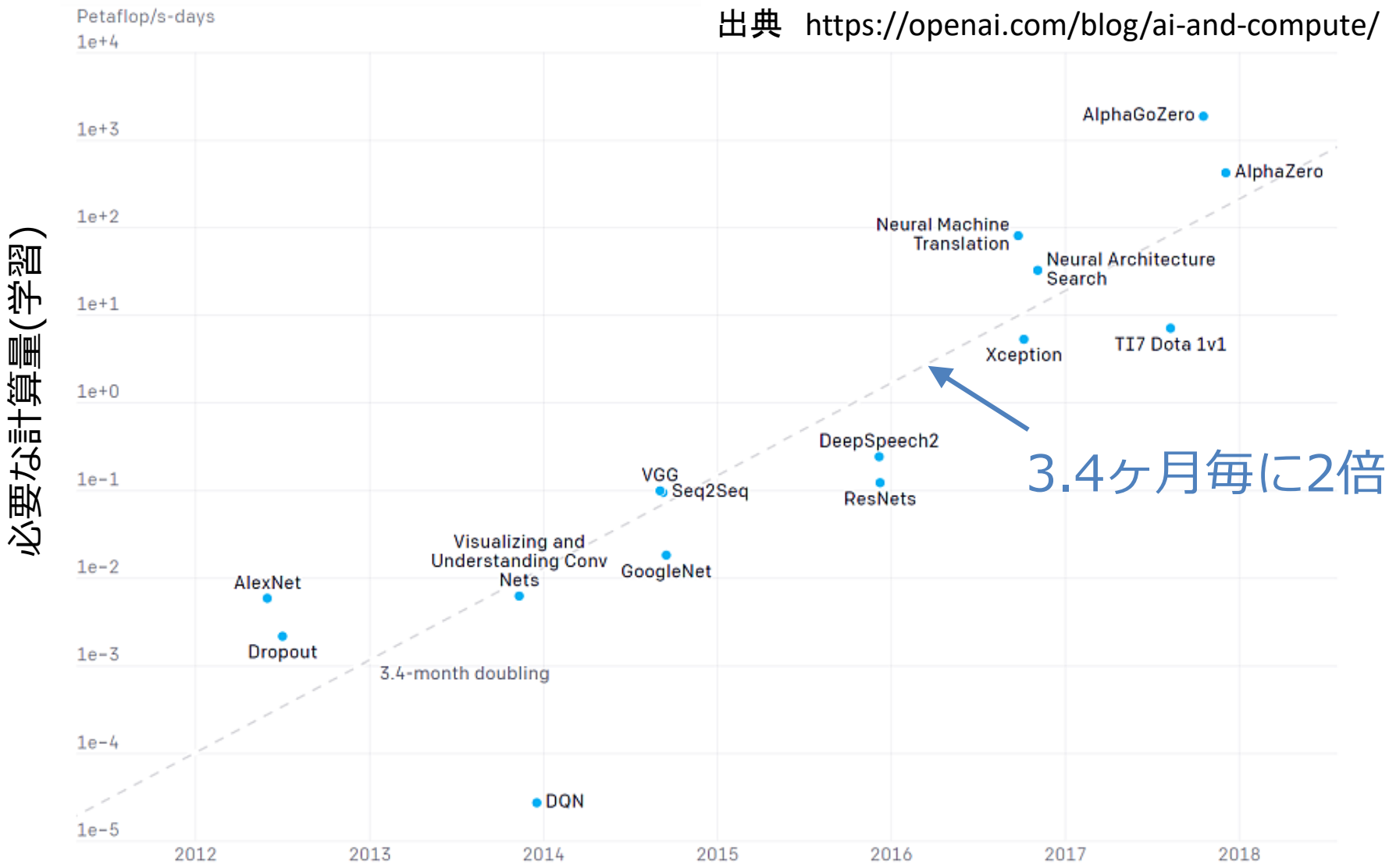
サービス事業者によるチップ開発

- **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**
  OoO RISC-V(Xuantie-910)を開発中 (2020/ISCA, Hot Chips)
  12nm, 12段パイプライン, ベクトル拡張命令, 2-2.5GHz
- **Preferred Networks**
  学習用AIチップ(MN-Core)を開発（2018, 2020/ISC)
  524TFLOPS, 500W/4die

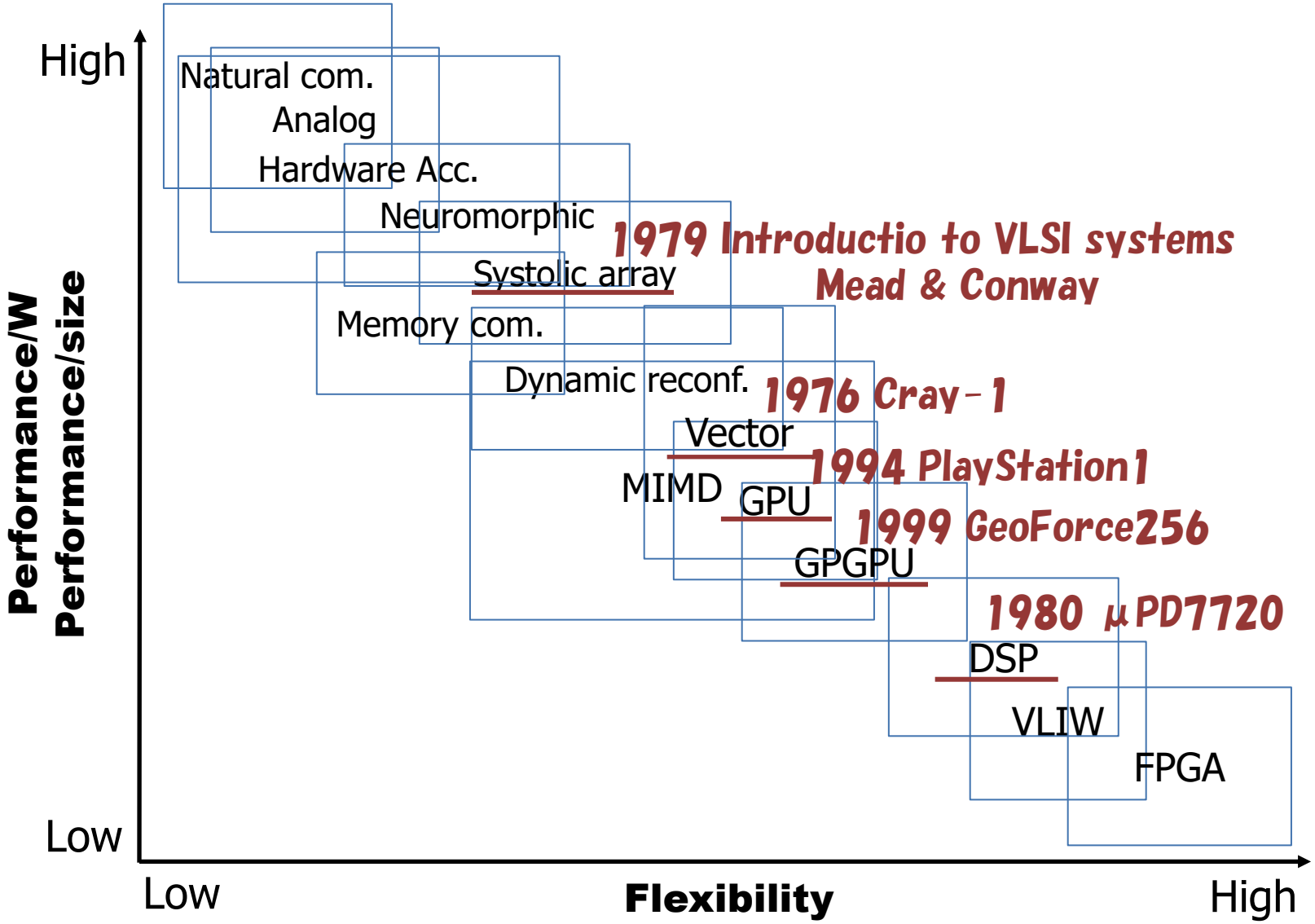
AI処理($y_i = \sum_j w_{ij} x_j$)の計算量

Peta operations (1000兆演算)

出典　https://openai.com/blog/ai-and-compute/



Domain Specific Hardware(アクセラレータ)のアーキテクチャ



AI関連チップ・ベンダー

Nvidia(US): GPU
Intel(US): Nervana NN processor, FPGA
Xilinx(US): FPGA
Wave Computing(US): 16,000 cores on chip, $117M
Cerebras Systems(US): DL training chip, $112M
Groq(US): AI training chip, 400TOPS, $10.3M
Mythic(US): Flash-type NN processor, $55M
Graphcore(UK): 1,000 cores on chip, $110M
Cambricon Technology(China): DL processor, $100M
Horizon Robotics(China): Embedded AI proc., $100M
DeePhi(China): CNN proc., $40M
DMP(JPN): Configurable AI inference processor IP
NSITEXE(JPN): Data flow processor
AXELL(JPN): DL chip
LeapMind(JPN): Inference accelerator IP
ArchiTeK(JPN): Edge AI processor
:

DARPA ERI(Electronics Resurgence Initiative) プロジェクト

- 2017年より開始されたDARPAによる半導体強化のための研究開発プロジェクト。**5年間で15億ドル**。
- **ムーアの法則の鈍化**と**中国の台頭**という2つの課題により、米国が半導体チップにおける競争力を失ってしまうことを懸念し、半導体産業の復興をかけ、米国のチップ開発・製造に革命をもたらす可能性のあるアプローチを探索する。
- 「**Materials & Integration**」、「**Architectures**」、「**Designs**」の 3 分野に投資。計19のプログラムを実施中。

米国における半導体強化施策（2020-2021年）

・2020年5月15日
TSMCが5nmファブ計画@アリゾナを発表

・2020年6月16日
CHIPS(Creating Helpful Incentives to Produce Semiconductors)**法案**を議会提出

・製造装置に対する投資税額控除(40%～)
・先端ファブへの100億ドル マッチング資金
・研究開発：120億ドル

・2021年3月23日
Intelが2棟のファブ計画@アリゾナを決定
ファウンドリサービスに参入

・2021年4月12日
バイデン大統領提案
半導体の製造と研究に500億ドル

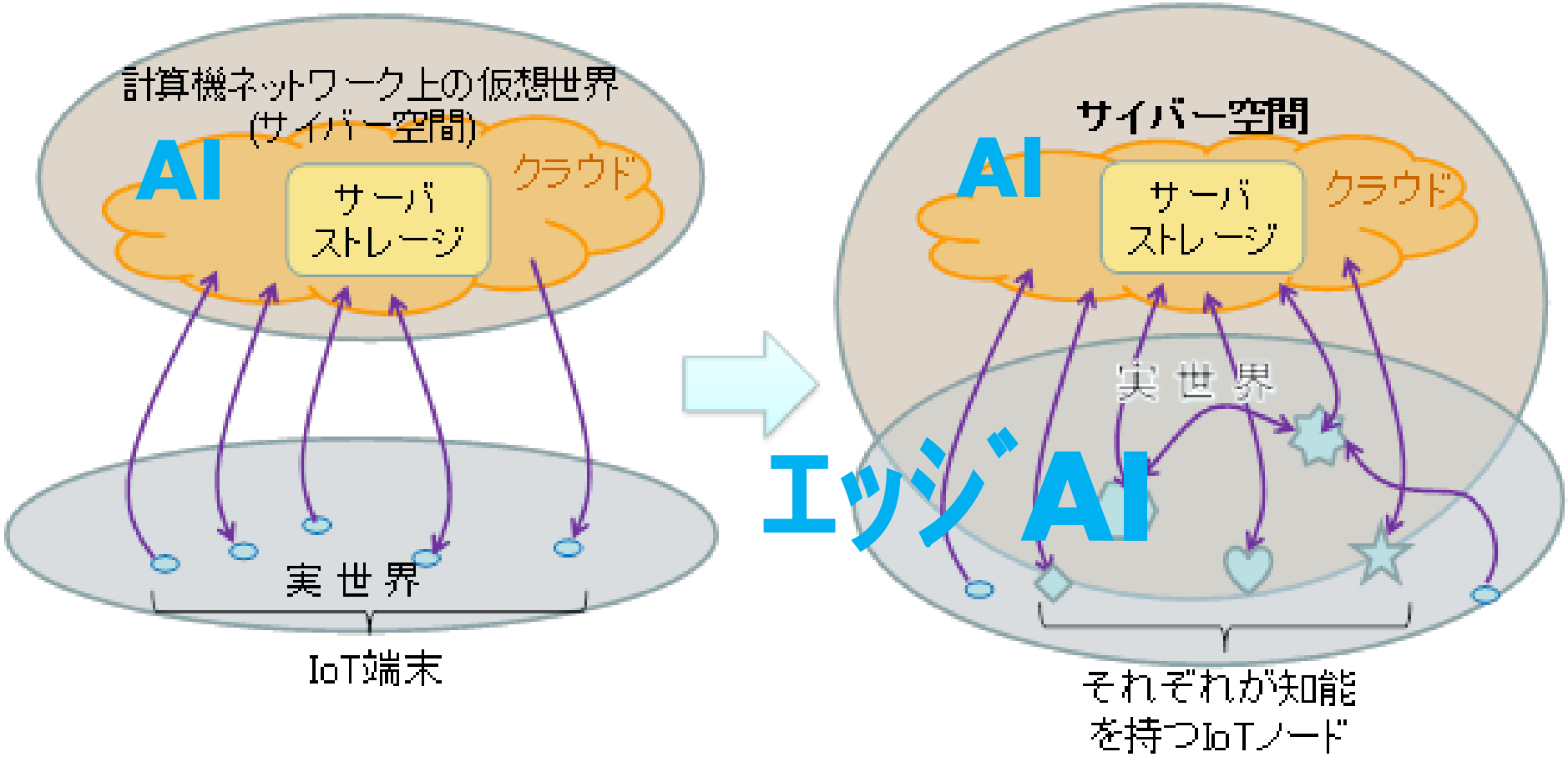
中国における半導体強化策

- **「中国製造 2025」**
  2015年５月発表の産業政策
  半導体自給率を 2025 年までに 70%に引き上げる

- **国家集積回路産業投資基金**
  中国政府主導の半導体産業育成のための巨大ファンド
  第１期：2014年、1387億元（約2兆円）
  77件のプロジェクトと55社の半導体メーカーに資金を提供（～2018年）
  第２期：2019年、2041.5億元（約3兆円）
  2019年末より投資スタート

2017年度EAJ政策提言プロジェクト
「次世代コンピューティング技術」

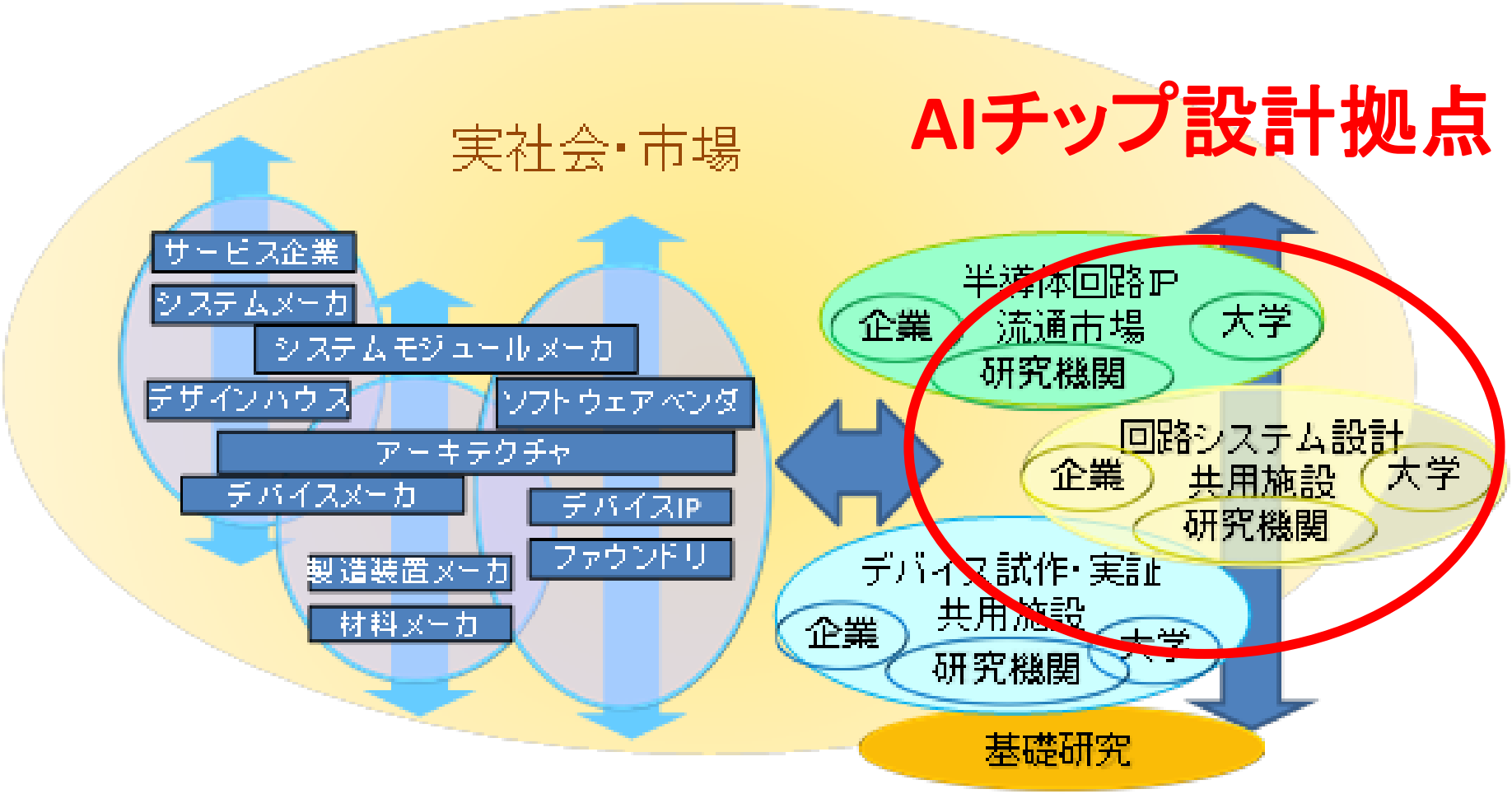
実世界とｻｲﾊﾞｰ空間でのﾃﾞｰﾀ処理を融合した事業ﾓﾃﾞﾙへの移行



出典：EAJ報告書「2030年に向けた次世代計算機技術開発戦略」

2017年度EAJ政策提言プロジェクト「次世代コンピューティング技術」

エコシステム型企業連携とLSIの開発実証を支援する共用設計・試作施設



出典：EAJ報告書「2030年に向けた次世代計算機技術開発戦略」

AIﾁｯﾌﾟ・次世代ｺﾝﾋﾟｭｰﾃｨﾝｸﾞ関連の施策

・**経産省**

AIチップ・次世代コンピューティングの技術開発事業
NEDO

AIチップ開発加速のためのイノベーション推進事業
NEDO

ポスト５G情報通信システム・半導体の開発
NEDO

・**文科省**

Sosiety5.0を支える革新的コンピューティング技術
JST/CREST

革新的コンピューティング技術の開拓
JST/さきがけ

２．AIチップ設計拠点

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」

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。

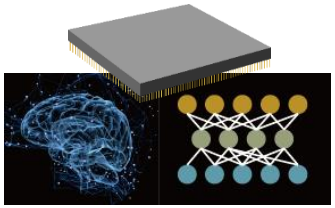
革新的AIチップのアイディア

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**

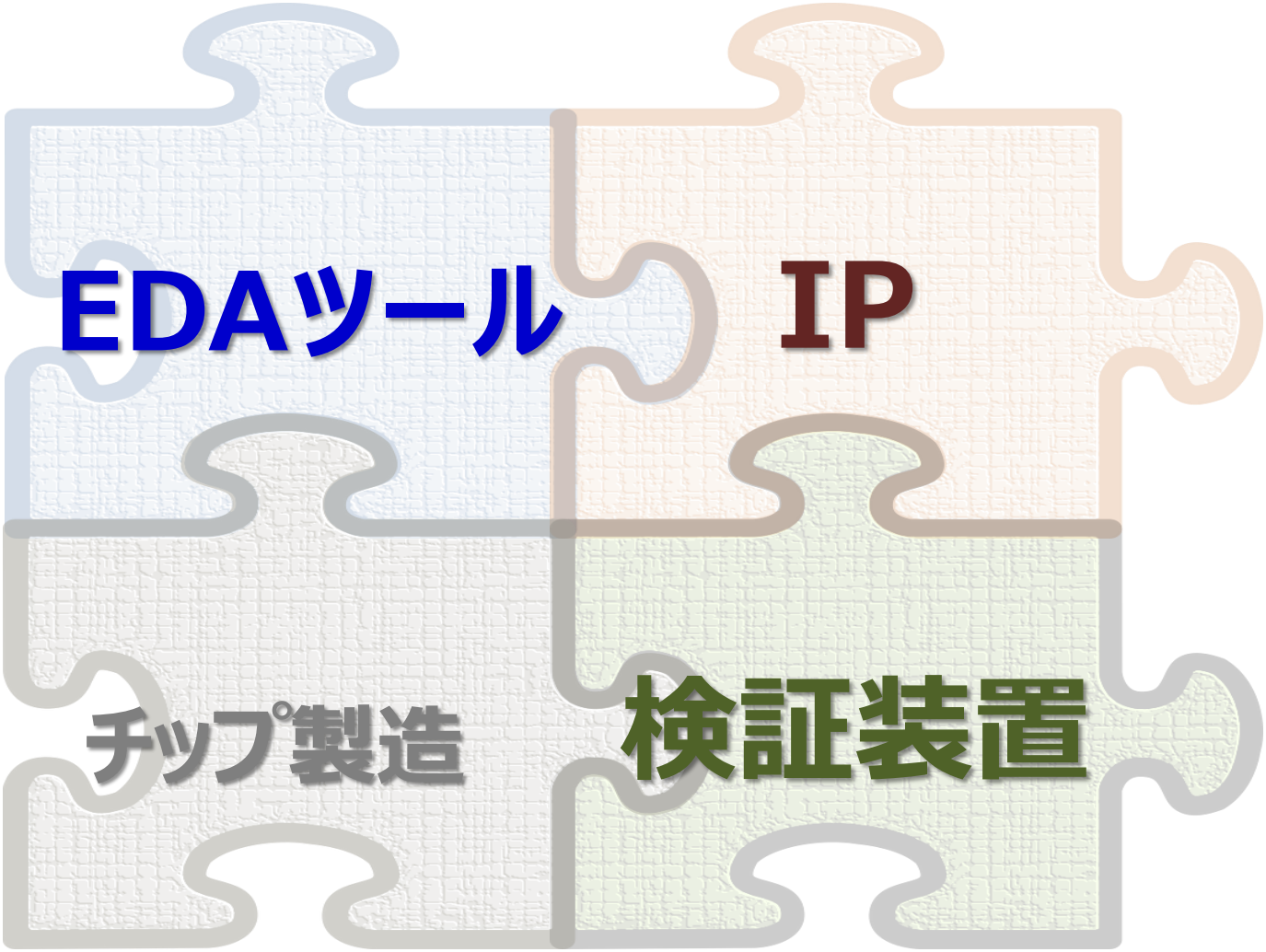
：

拠点の体制・運営

赤字：AIチップ設計拠点 運営組織

- ✓ 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

拠点での整備



EDAツール

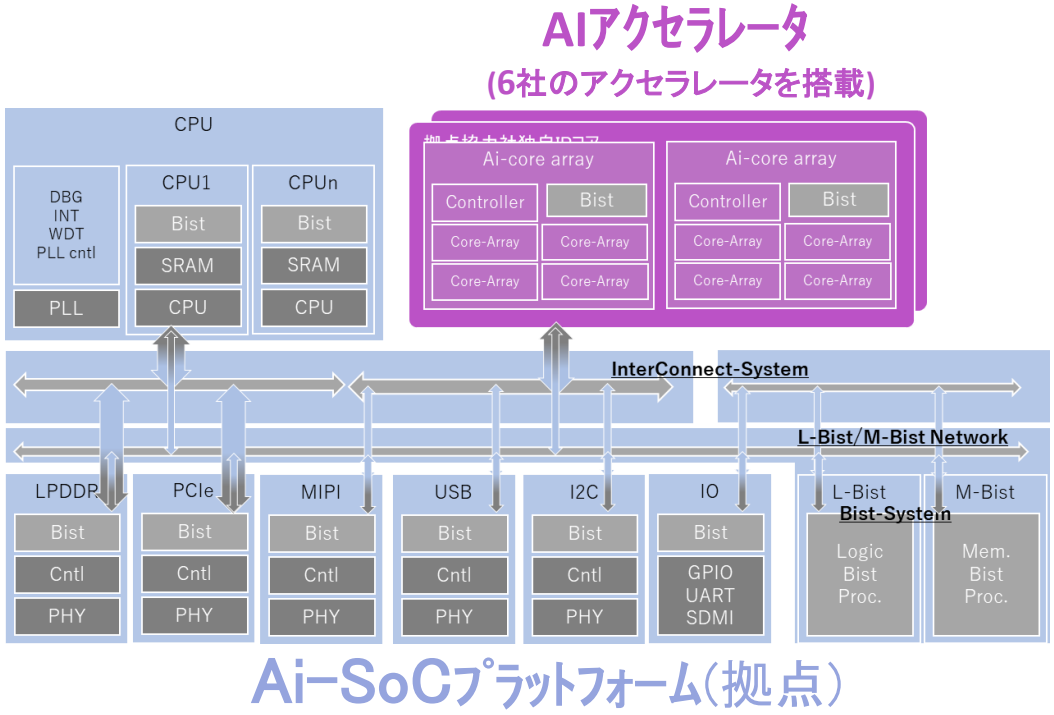
Cadence, Synopsys, Mentorのツールを整備

IP

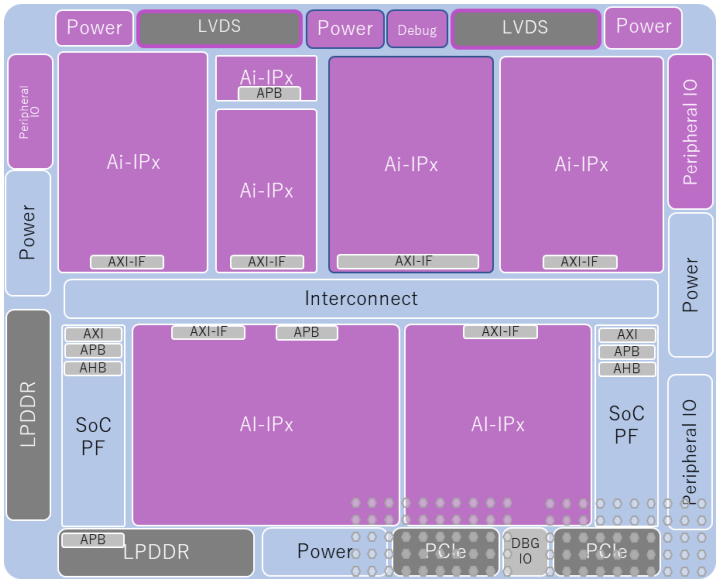
Synopsysの40nm, 28nm IPパッケージを整備
（物理系はTSMC向け）

乗合チップ (Ai-One)

- 拠点導入IP(28nm)を活用して、拠点がSoC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、実証実験を進める



チップ内部構成

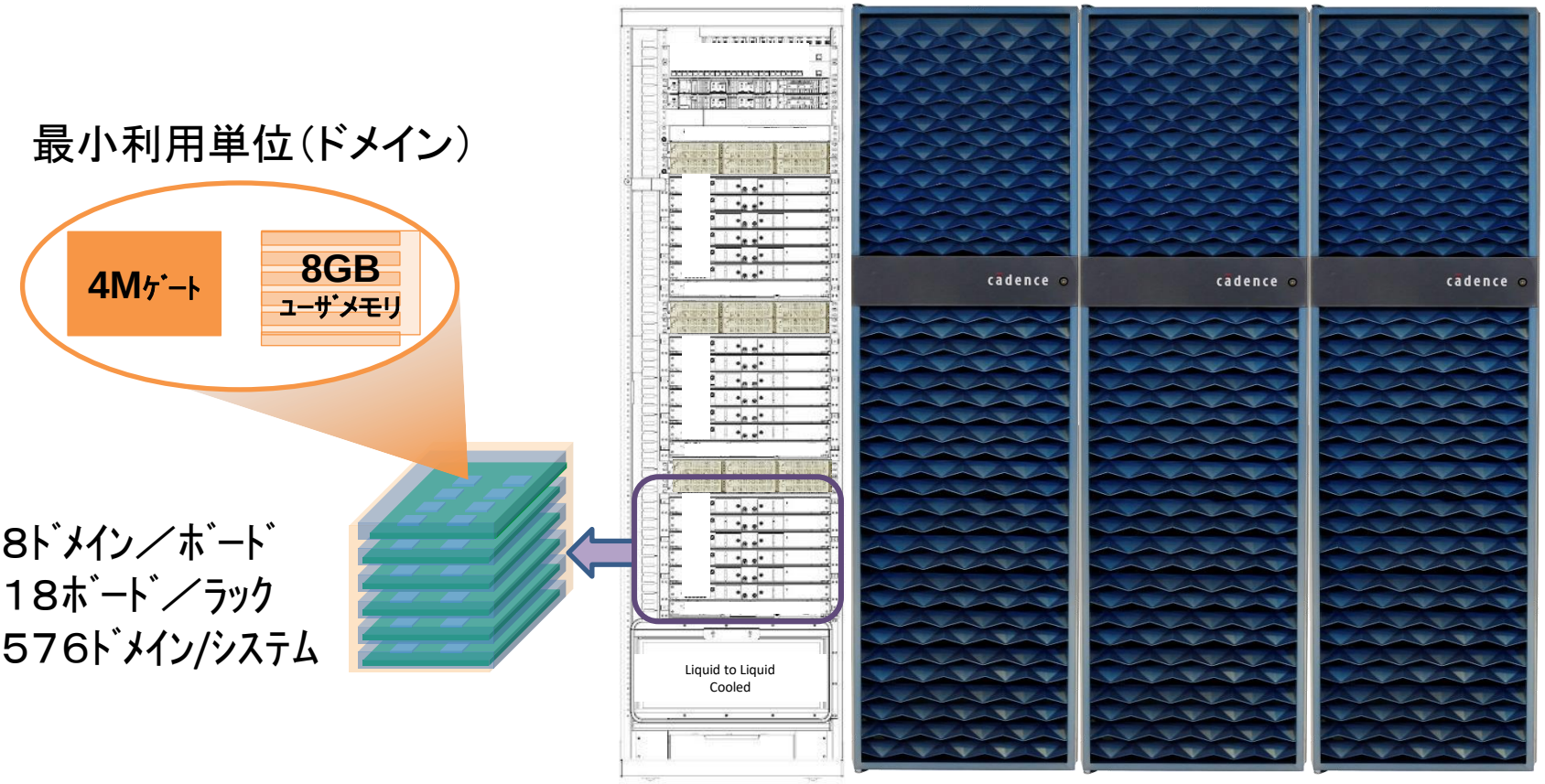


チップ実装イメージ

論理エミュレータの概要

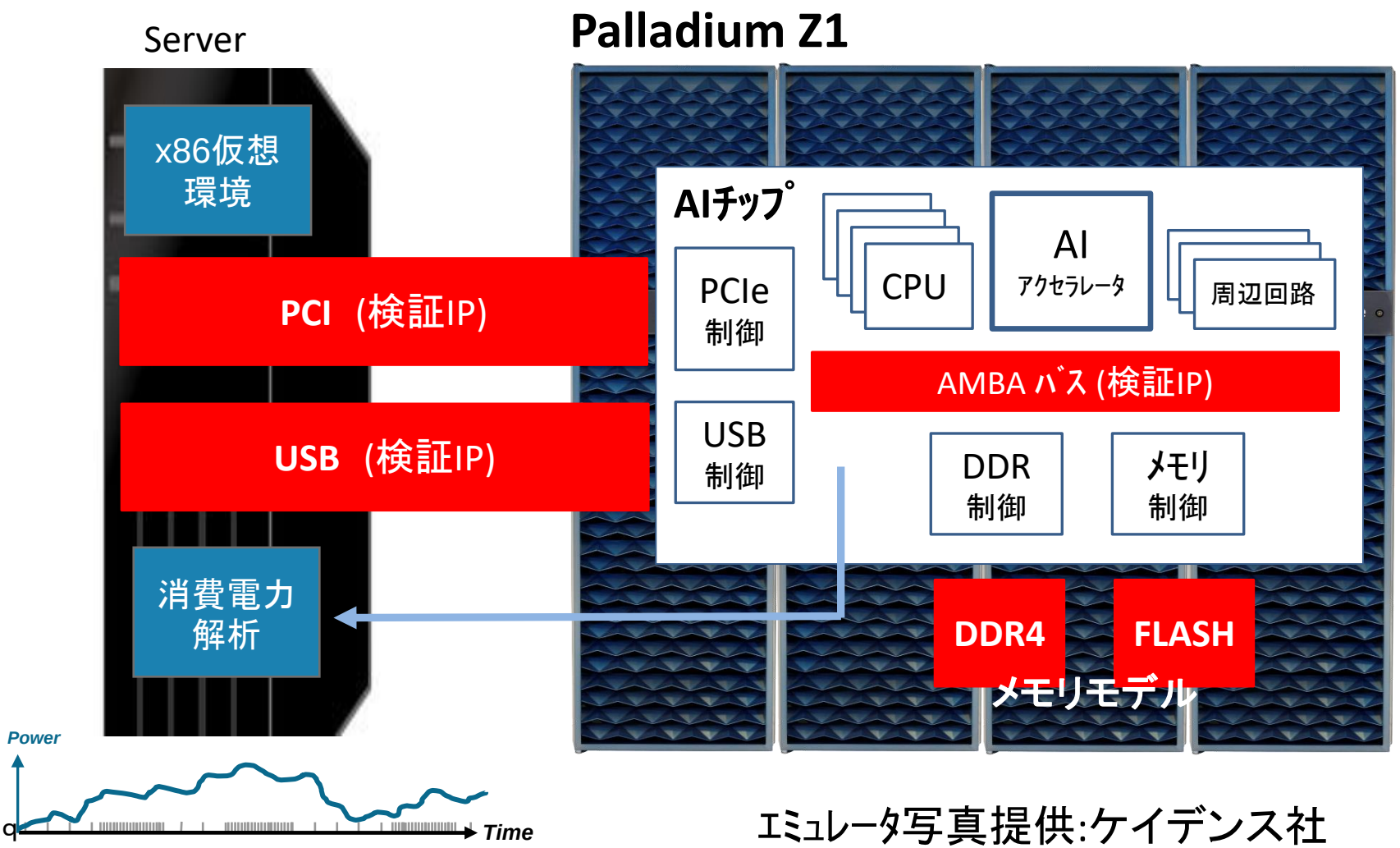
Palladium Z1

- 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



エミュレータ写真提供:ケイデンス社

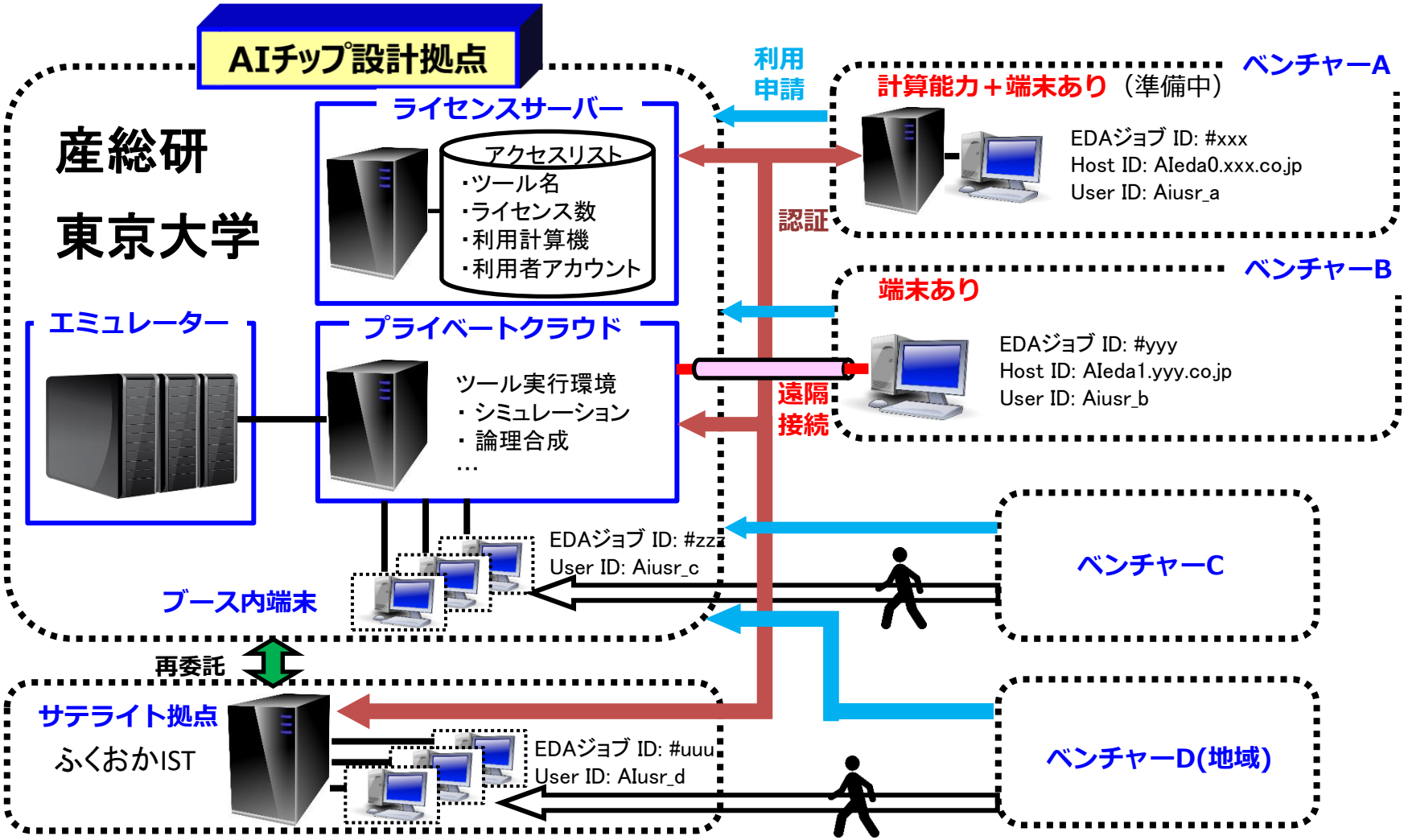
論理エミュレータの活用例



エミュレータ写真提供:ケイデンス社

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18回 AIチップ設計拠点フォーラム（12/23）

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　**オンデバイス学習とそのチップ化に向けて**
（慶應義塾大学　松谷宏紀先生）

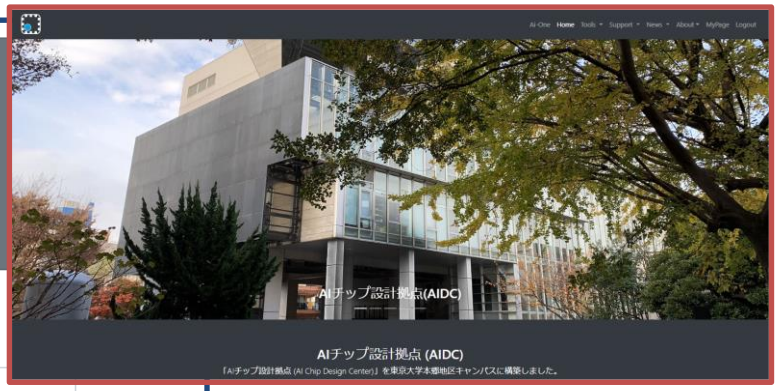
14:35-15:35　**IEDM2020に見る半導体技術研究動向**
（産総研　五十嵐泰史氏、更田裕司氏）

15:40-16:40　**AI at the Edge – Making it Work**
（VeriSilicon　Simon Jones氏）

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

プロジェクトID申請には、

既にプロジェクトIDをお持ちで追加機能申請の場合は、本フォームから申請ください。
プロジェクトに参加される拠点利用者全員のEmailアドレスを記載願います。
プロジェクトID入手後、拠点利用者全員に拠点ID申請をお願いします。
(プロジェクトID申請は、管理責任者以外は必要ありません)



プロジェクト

プロジェクトID(1)	新規作成の場合は、ブランクです。
プロジェクト名(1) 必須	任意のプロジェクト名を記載ください。
プロジェクト概要(1) 必須	研究開発概要、開発期間、使用ツール、プロセス、
プロジェクトID(2)	新規作成の場合は、ブランクです。
プロジェクト名(2)	任意のプロジェクト名を記載ください。
プロジェクト概要(2)	研究開発概要、開発期間、使用ツール、プロセス、
お名前 確認	内山
ミドルネーム 確認	ミドルネーム（例：Joanne Angelina ）
Emailアドレス 確認	kunio.uchiyama@aist.go.jp
電話番号 確認	09093784330
所属機関 確認	産業技術総合研究所

AIチップ設計拠点の装置等利用規約に合意

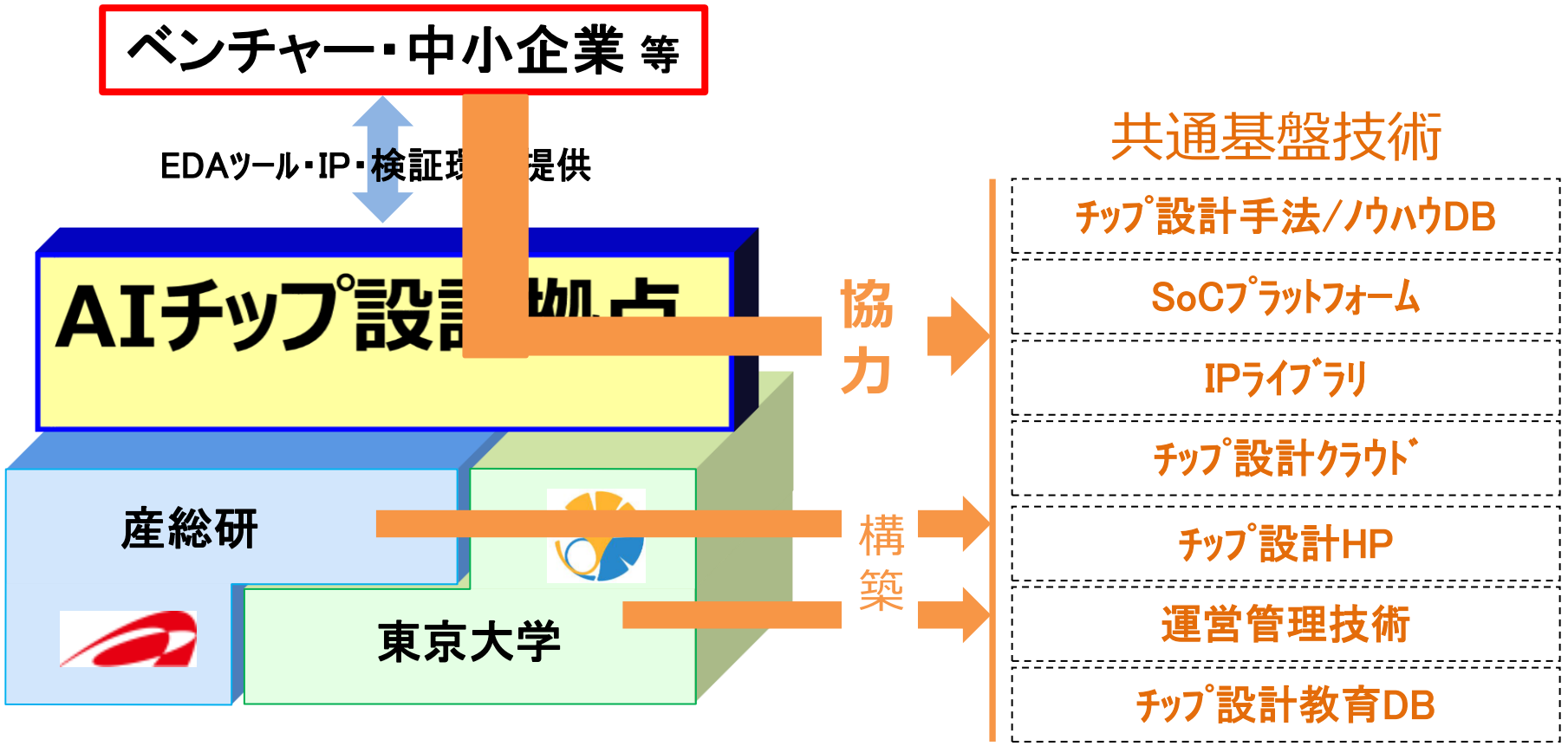
問い合わせ

お名前 必須	姓 (例：山田) / 名 (例:太郎)
お名前（カタカナ フリガナ） 必須	姓（例：ヤマダ） / 名（例：タロウ）
会社名（法人の方） 必須	所属機関（例：株式会社○○○○、○○大学）派遣元や複数所属している機関も全て記載ねがいます。
メールアドレス 必須	例：yourname@sample.com
お問い合わせ内容 必須	お問い合わせ内容をご記載ください。

設計拠点のプライバシーポリシーに ☐ 同意する 必須

送信する

共通基盤技術の構築


ベンチャー・中小企業 等
EDAツール・IP・検証
提供
AIチップ設計拠点
産総研
東京大学
協力
構築
共通基盤技術
チップ設計手法/ノウハウDB
SoCプラットフォーム
IPライブラリ
チップ設計クラウド
チップ設計HP
運営管理技術
チップ設計教育DB

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

ご清聴ありがとうございました。