資料2-3

理想とする10年後の社会を実現するために、今、何を研究し、どのような人材を育成すべきか

令和6年２月16日

東京大学 工学系研究科 教授

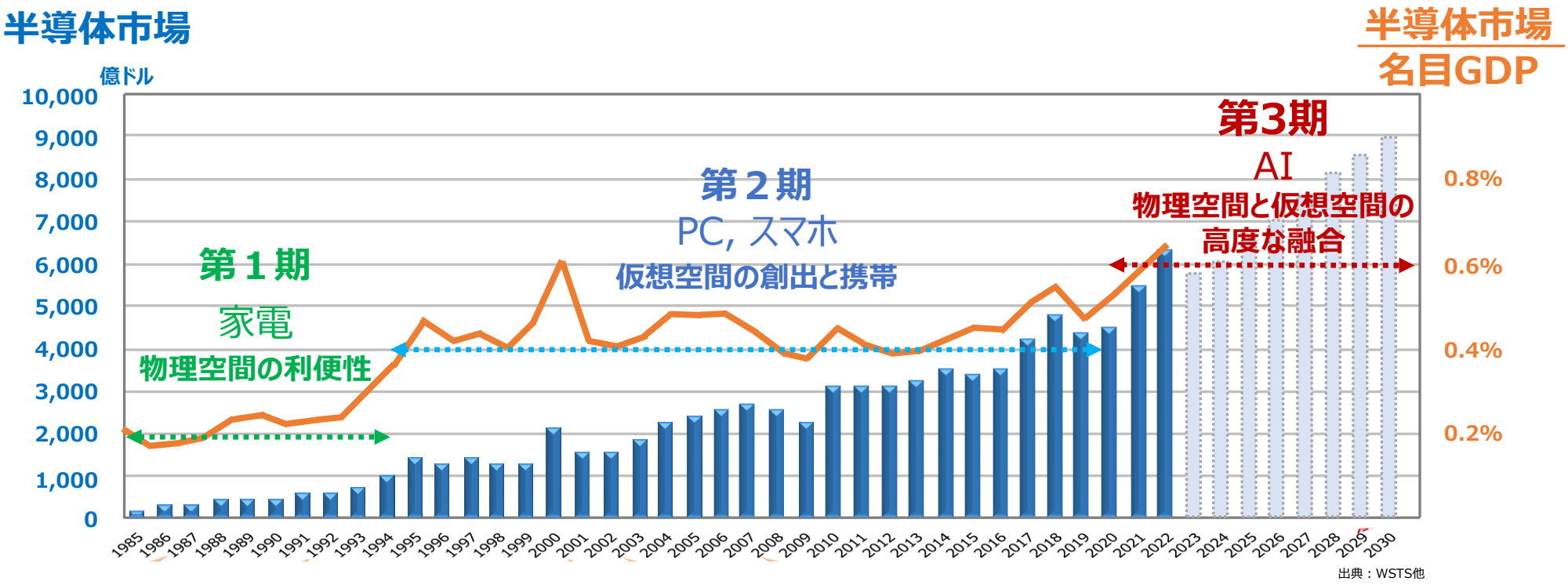
東京大学 THE UNIVERSITY OF TOKYO

黒田　忠広

AIが半導体需要を生むと共にエネルギー危機を招く

- 第3期成長期を迎え、物理空間と仮想空間の高度な融合で価値を創る
- AIが半導体の需要を創出し、2030年には市場の70%(760B$)がAI半導体になる(1)
- 2030年には現在の総電力の2倍、2050年には200倍もの電力をIT機器だけで消費(2)
- 大型火力発電換算で2030年に28基、2050年に4,500基が必要(3)

出典 (1) IBS Inc.
(2) 経済産業省『半導体・デジタル産業戦略』
(3) 経済産業省第5回半導体・デジタル産業戦略からSoftBankが試算



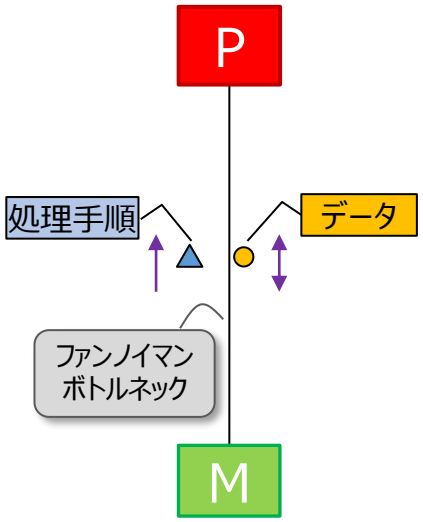
出典：WSTS他

AI処理のボトルネックはGPUではなくDRAM

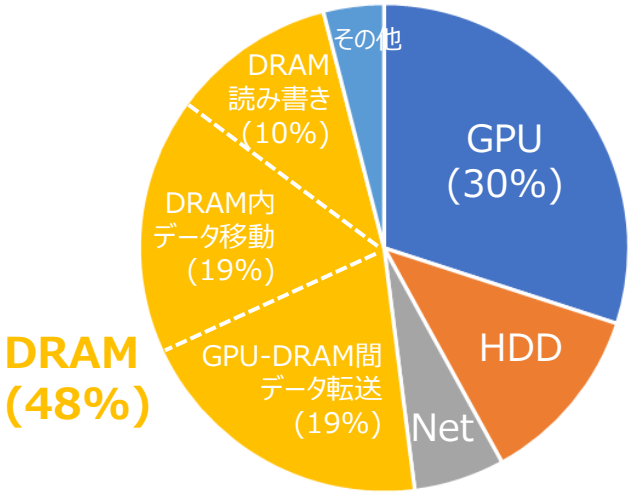
- フォンノイマンボトルネック(DRAMアクセス)で大半の電力を消費(1)
- NVIDIAのH200はGPU同一でもメモリの帯域1.4倍・容量1.7倍で推論性能1.6倍(2)
- HPCの多くのアプリでもメモリ帯域が性能を律速(3)

出典 (1) *ACM Tras. Architecture and Code Opt*.vol. 10, Issue 4, pp. 1-25, 2013
(2) https://doi.org/10.1007/978-3-030-80126-7_35
(3) https://www.hpcwire.jp/archives/80667

フォンノイマンボトルネック

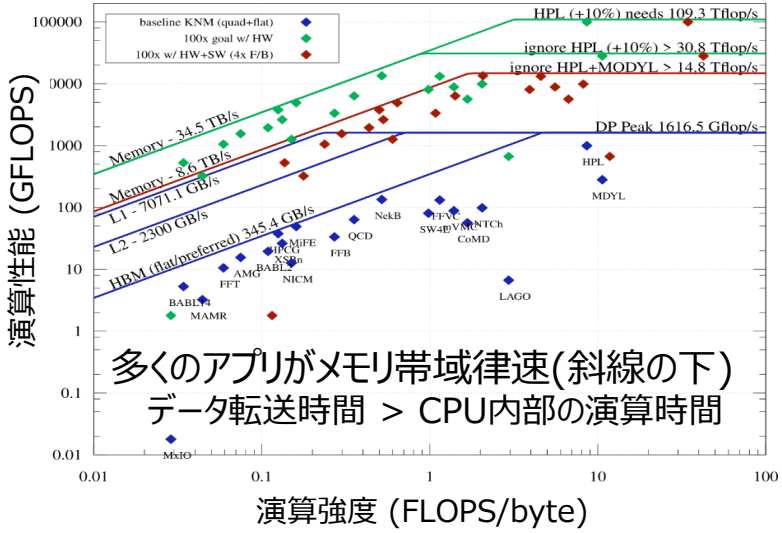


AI時代のGPUクラウドサーバー



J. Zhao *et al.*, "Optimizing GPU energy efficiency with 3D die-stacking graphics memory and reconfigurable memory interface," in *ACM Tras. Architecture and Code Opt*.vol. 10, Issue 4, pp. 1-25, 2013. 本論文を基にAI推論の電力内訳を分析

A64FX上でのルーフライン解析



Yang, C., Wang, Y., Kurth, T., Farrell, S., Williams, S. (2021). Hierarchical Roofline Performance Analysis for Deep Learning Applications. In: Arai, K. (eds) Intelligent Computing. Lecture Notes in Networks and Systems, vol 284. Springer, Cham.

演算強度とは、メモリから読みだした1バイトのデータを使ってプロセッサが何回演算できるかを示した値。演算強度が高い領域では、メモリから1回データを読み出せば何回も計算ができるので、演算性能が計算システムの性能律速になり、屋根の高さが直線状になる。
演算強度が低い領域では、計算のたびに何度も新しいデータをメモリから読み出す必要があるので、メモリ帯域が計算システムの性能律速になり、演算強度に比例した右上がりの直線となる。AI処理は、処理が進むごとに新しい重み係数を読み出すことが必要なため、演算強度が小さいタスクである。

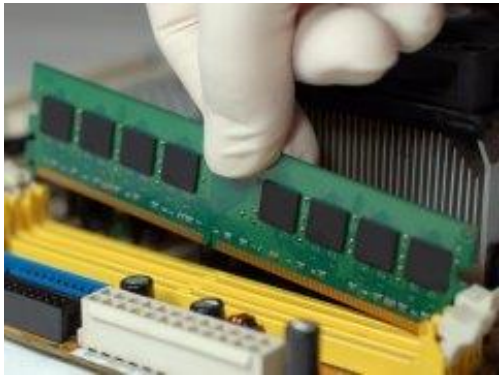
NVIDIA GPU
H100(2023/3) ：HBM2e 3.4TB/s, 80GB
H200(2023/11)：HBM3e 4.8TB/s, 141GB

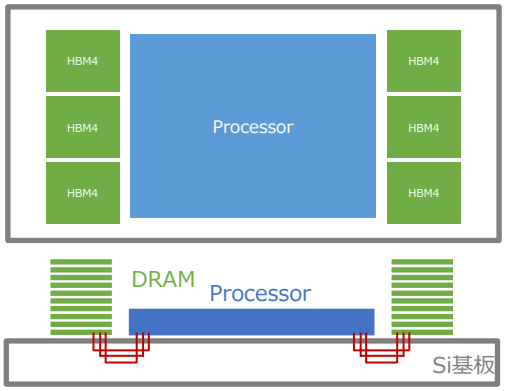
3D集積でボトルネック解消、日本に学術の蓄積あり

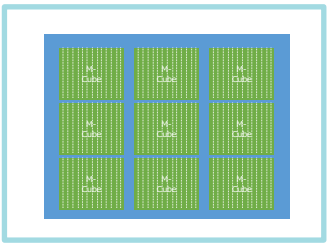
- 2.5D->3Dでボトルネック解消
- 3Dの課題は抜熱、材料と製造装置が実用化の鍵、日本に一日の長あり(1)

備考 (1) TSMCやSamsungが日本に3DIC研究開発拠点を開設

DIMM → **HBM 2.5D** → **3D**

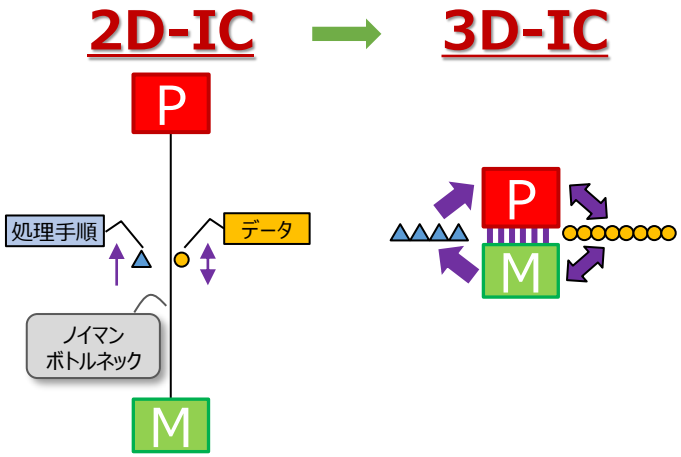


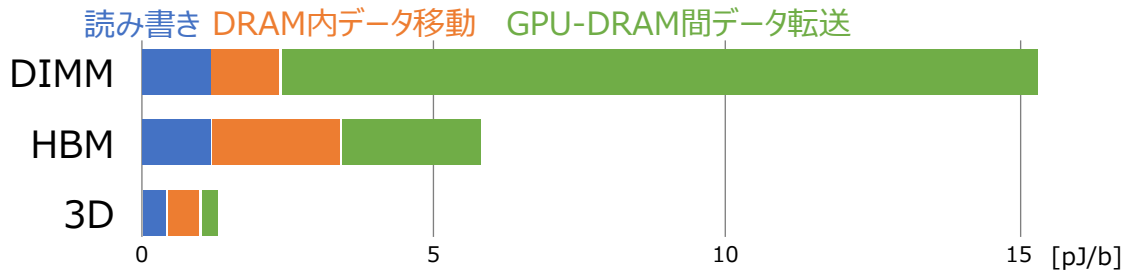




課題は抜熱

~15.3 pJ/bit (2.7)　　~5.7 pJ/bit (1)　　~1.3 pJ/bit (**1/4.4**)





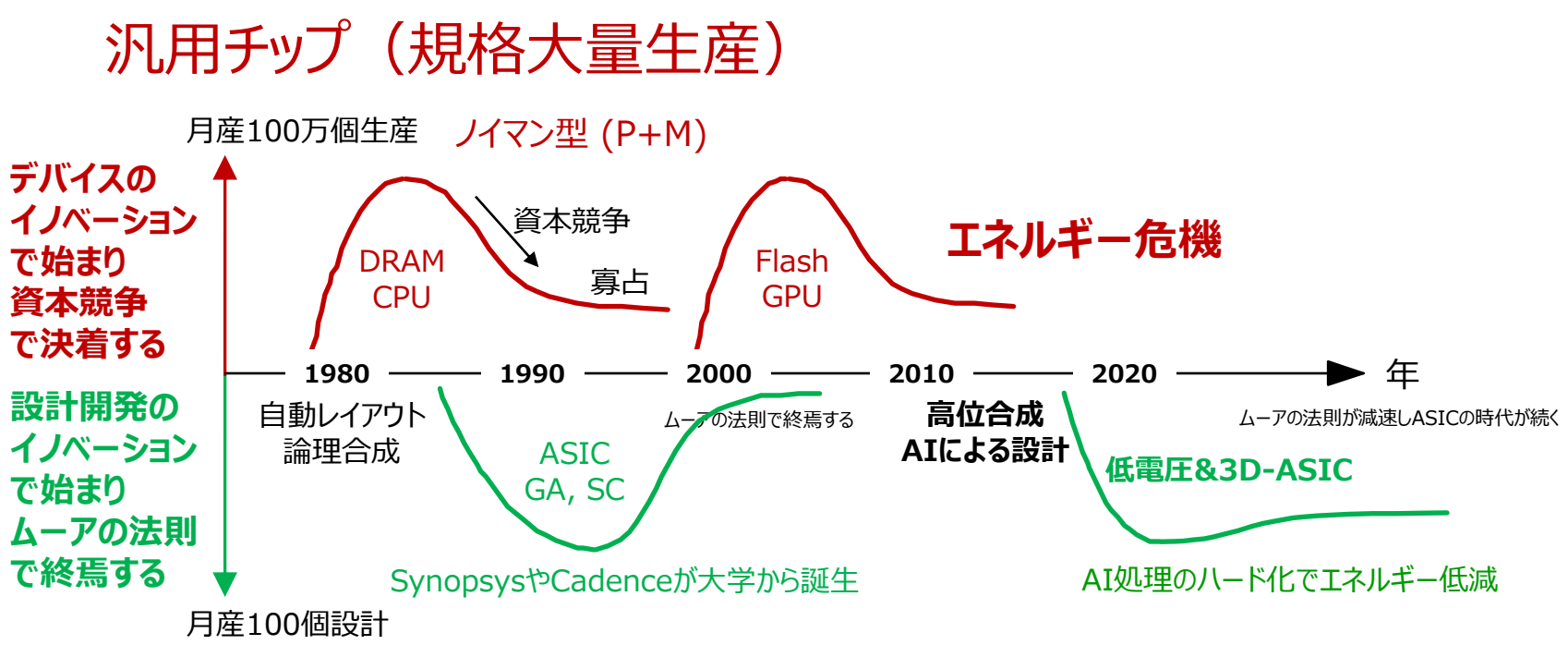
DRAMの単位ビット当たりの消費エネルギー

[Ref] M. O'Connor (NVIDIA),
"Fine-Grained DRAM: Energy-Efficient DRAM for Extreme Bandwidth Systems,"
IEEE/ACM International Symposium on Microarchitecture, 2017.

エネルギー効率の高い専用チップを効率良く開発する

- 汎用チップ(資本競争)の時代から専用チップ(設計競争)の時代に移る
- 低電圧設計＆3D集積を可能にするDMCO(1)が競争力の源泉

備考 (1) DMCO(*Design Manufacturing Co-Optimization*)はRapidusの戦略、設計と製造の協調最適化。
エネルギー消費は電源電圧の２乗に比例するので電源電圧を下げることが有効。しかし、電源電圧を下げるとしきい値電圧のばらつきが性能に大きく影響する。低電圧設計を可能にするプロセス情報(PDK)と3D集積を可能にするパッケージ情報(PADK)を設計者に提供することが重要。



出典：T. Kuroda, ISSCC 2010 Panel Discussion, “Semiconductor Industry in 2025”.

AIを駆使した自動設計、国内のEDA人材は絶滅危惧種

- 設計コストが急増、コンピュータによる自動設計が必須 (No human in the loop)
- 国内のEDA人材は絶滅危惧種、人材育成が急務

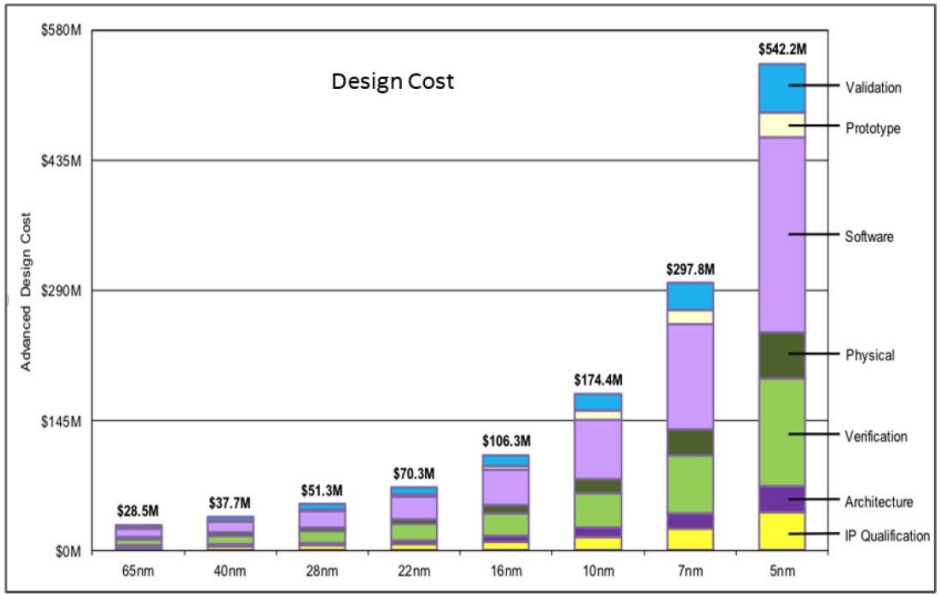
専用チップの開発
5nm試作の場合

200人
2年, 800億円

ファウンドリ
4ヶ月, 20億円

設計

製造

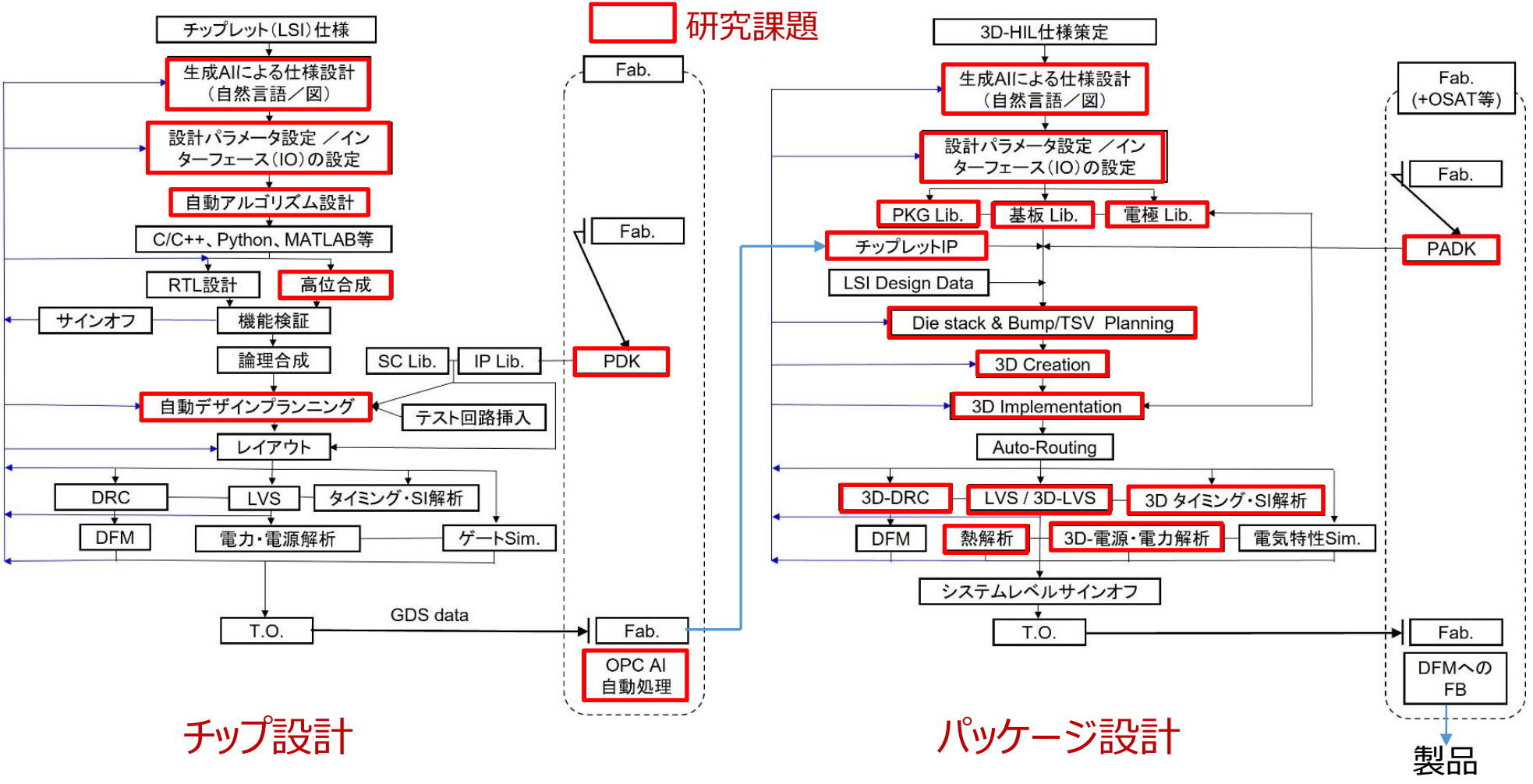


SoC設計コストは微細化と共に飛躍的に増えている
$542.2M@5nm

出典：UCIe1.1_White_Paper_2023_FINAL

3DICの設計フロー

- チップ設計と同程度にパッケージ設計が複雑になる
- EDAの研究課題が山積、すべて米国に依存している点が日本の課題

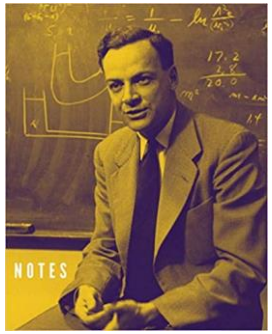


ユースケース創出は俯瞰力(10^{-18}の物理～10^{18}の情報)

- 一兆個のトランジスタで何を作るか？HWとSWのリソースをどのように配分するか？
- 広大な空間(物理と情報、デバイスとシステム、HWとSW)を俯瞰する力が求められる

There's Plenty of Room at the Top.

演算性能 E (10^{18})
通信性能 P (10^{15})
ギガ集積 T (10^{12})
集積度 G (10^{9})
メガ集積 M (10^{6})
抵抗 k (10^{3})
電力 (1)
電圧 m (10^{-3})
電流 μ (10^{-6})
寸法 n (10^{-9})
エネルギー p (10^{-12})
伝搬時間 f (10^{-15})
静電容量 a (10^{-18})
マイクロエレクトロニクス
ナノエレクトロニクス



物理と情報
デバイスとシステム
ハードとソフト
全体を俯瞰する力

There's Plenty of Room at the Bottom.
(by Richard P. Feynman, December 29th, 1959)

民主化がイノベーションを誘発、大学が社会インフラになる

- アイデアの交配でイノベーションは生まれる、頭脳が集まる大学が民主化拠点となる
- Serendipityではなく、技術・市場の将来予測から正しくバックキャストする能力が必要

■ AIの民主化

ChatGPT

専門家から一般人へ

■ 半導体の民主化

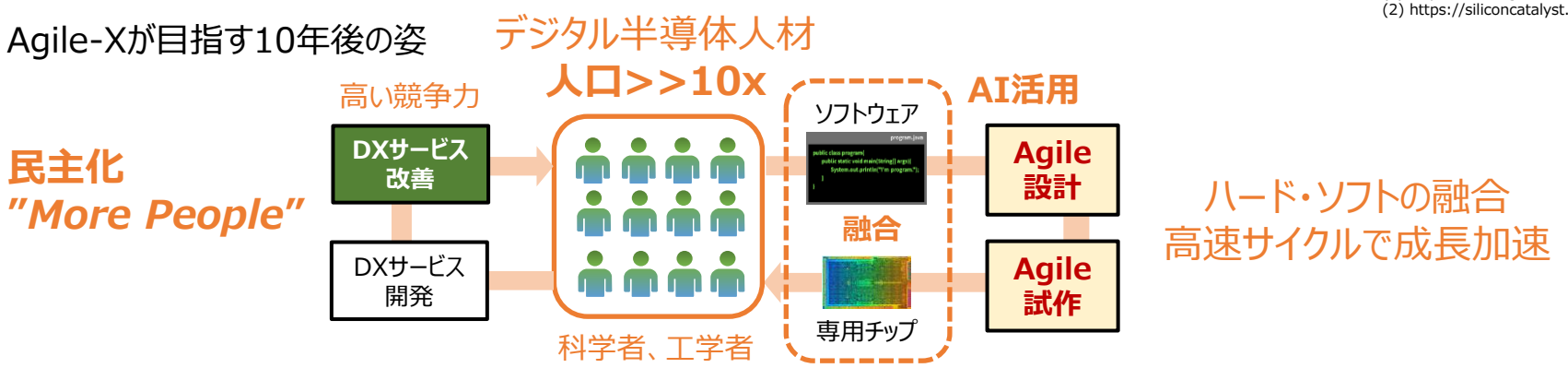
Innovation relies on the free flow of ideas, and ideas come from people
– a global alliance of innovators, facilitated by an open innovation platform.
ISSCC 2021, Mark Liu

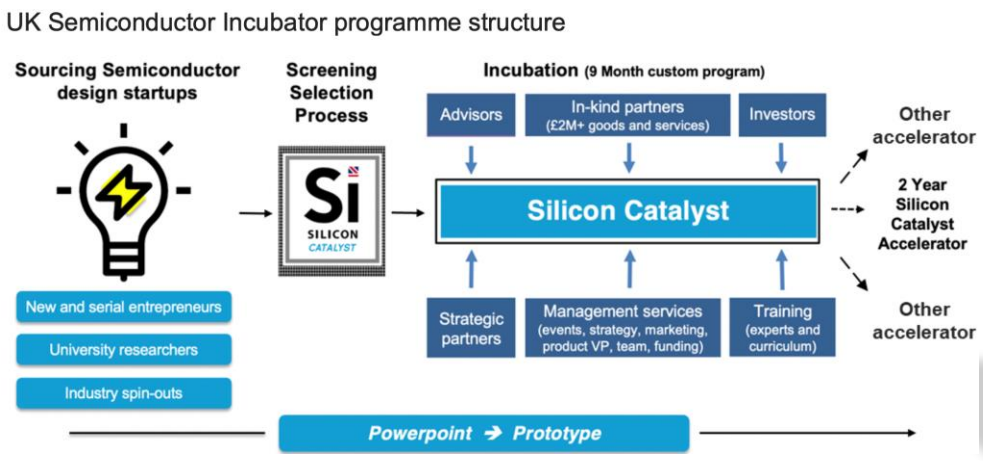
シリコン for サイエンス, エンジニアリング, イノベーション

- 科学と工学に貢献するAgile-Xプラットフォーム(1)
- イノベーションに貢献するSilicon Catalyst(2)

備考 (1) http://www.agile-x.t.u-tokyo.ac.jp/
(2) https://siliconcatalyst.com/

Agile-Xが目指す10年後の姿





WE ACCELERATE SEMICONDUCTOR STARTUPS FROM POWERPOINT TO PROTOTYPE

まとめ：何を研究し、どのような人材を育成すべきか

【研究】

- エネルギー効率改善
  - 2.5Dから3Dへ
  - 低電圧設計＆3D集積のDMCO
- 開発効率改善
  - 汎用チップから専用チップへ
  - SWを書くようにチップを設計する

【教育】

- 俯瞰力
  - 10^{-18}の物理～10^{18}の情報
  - EDA人材を育てる
- 民主化
  - アイデアが交差する大学は社会インフラ
  - 設計プラットフォームを整備する

【多様性】

- 女性の参画
  - 技術系新入社員の女性比率：日系 11%, TSMC 21%, ﾏｲｸﾛﾝ･ｼﾞｬﾊﾟﾝ 43%
- 世界の頭脳
  - JST･ASPIREで国際頭脳循環を加速、Agile-Xで世界の頭脳を惹きつける