ムーアの法則による高速インターフェース展開予測2025/2026

Shishio Tsuchiya
shtsuchi@arista.com

ARISTA

Andy Bechtolsheim

https://en.wikipedia.org/wiki/Andy_Bechtolsheim

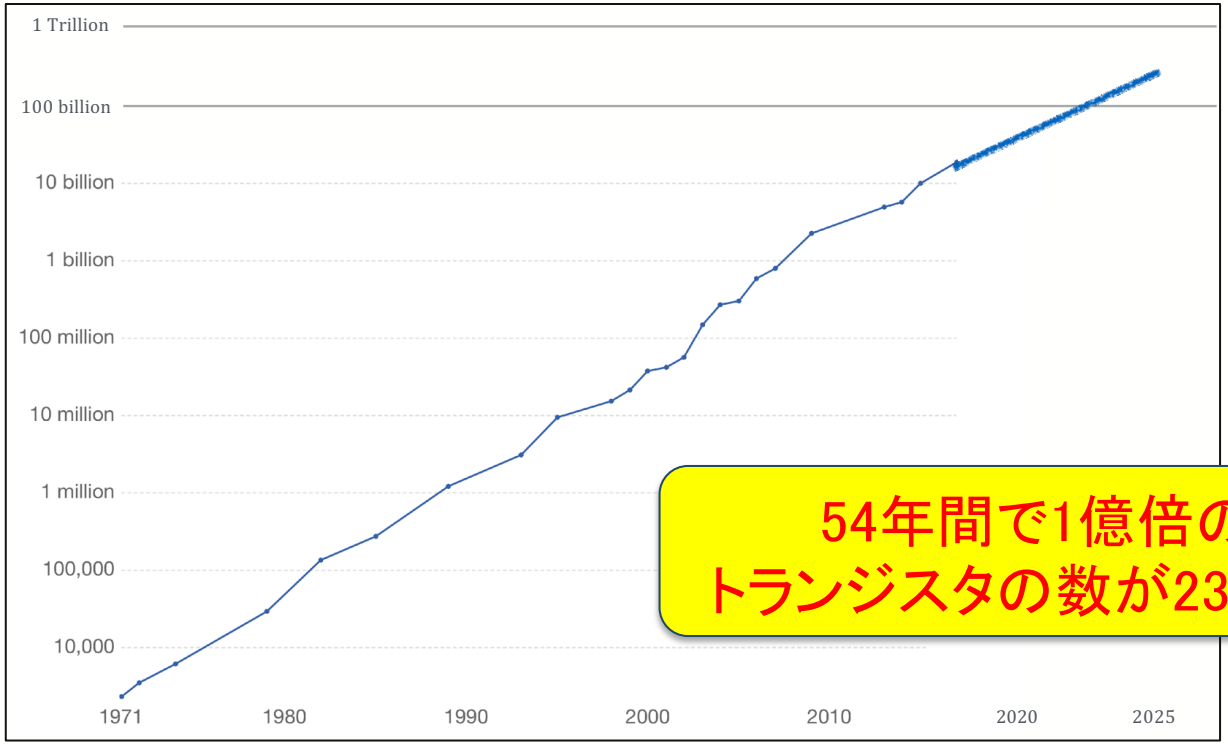
- サン・マイクロシステムズの共同創業者
- Granite Systems設立 ギガビットイーサネットを開発
  - 1996年にシスコシステムズに売却
- 1998年にラリー・ペイジとセルゲイ・ブリンに10万ドルの小切手を渡し、グーグルの最初の出資者に
- 2001年Opteron搭載サーバーのKealiaを設立
  - 2004年サン・マイクロシステムズに売却
- 2008年アリスタネットワークス創業

ムーアの法則

- ムーアの法則（ムーアのほうそく、英: Moore's law）とは、大規模集積回路（LSI IC）の製造・生産における長期傾向について論じた1つの指標であり、経験則に類する将来予測である。発表当時フェアチャイルドセミコンダクターに所属しており後に米インテル社の創業者のひとりとなるゴードン・ムーアが1965年に自らの論文上に示したのが最初であり、その後、関連産業界を中心に広まった。
- 彼は1965年に、集積回路あたりの部品数が毎年2倍になると予測し、この成長率は少なくともあと10年は続くと予測した。1975年には、次の10年を見据えて、2年ごとに2倍になるという予測に修正した。彼の予測は1975年以降も維持され、それ以来「法則」として知られるようになった。

https://ja.wikipedia.org/wiki/ムーアの法則

ムーアの法則 1971年～2025年 2年毎に2倍



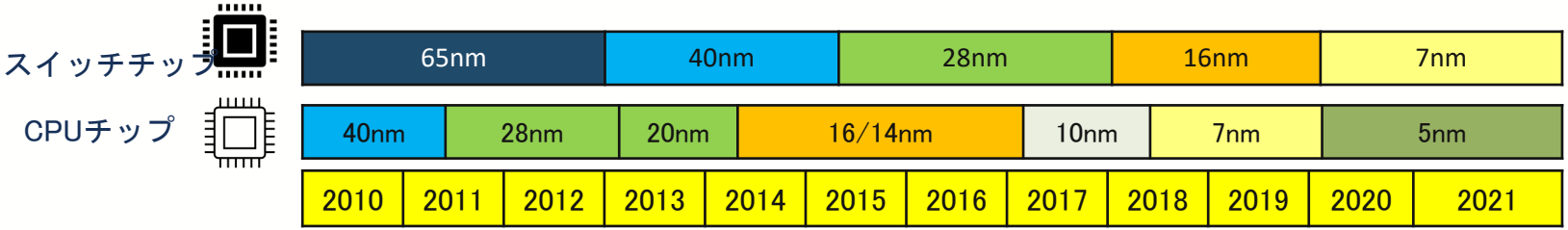
54年間で1億倍の改善(2^27)
トランジスタの数が2300から2500億へ

プロセッサー技術の向上

プロセスノード	7nm	5nm	3nm
相対密度	1	1.5	2.25
Speed@IsoPower	1	1.15	1.4
Power@IsoSpeed	1	0.8	0.6
量産体制	2019	2021	2023

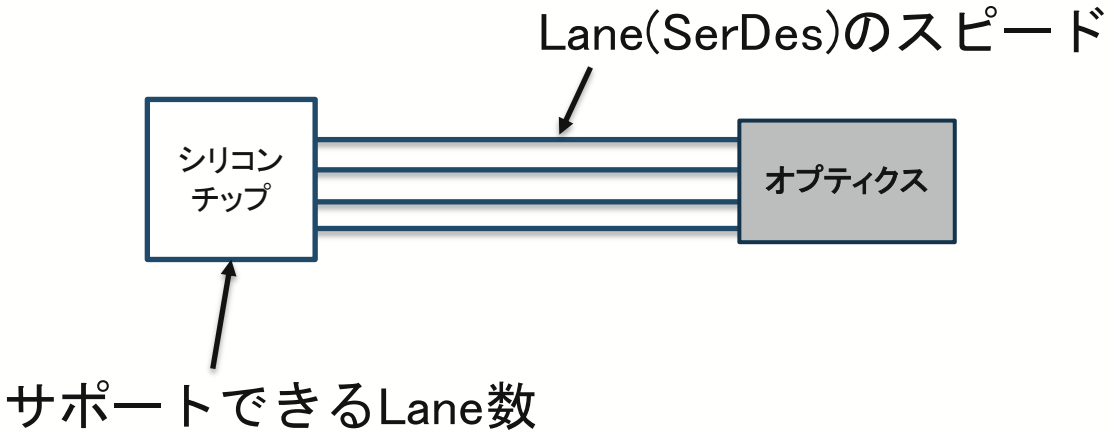
各プロセス世代で、スループットの向上、電力効率の改善、バッファの増加、ルーティングテーブルの拡大などが可能に

半導体技術の進化とロードマップ



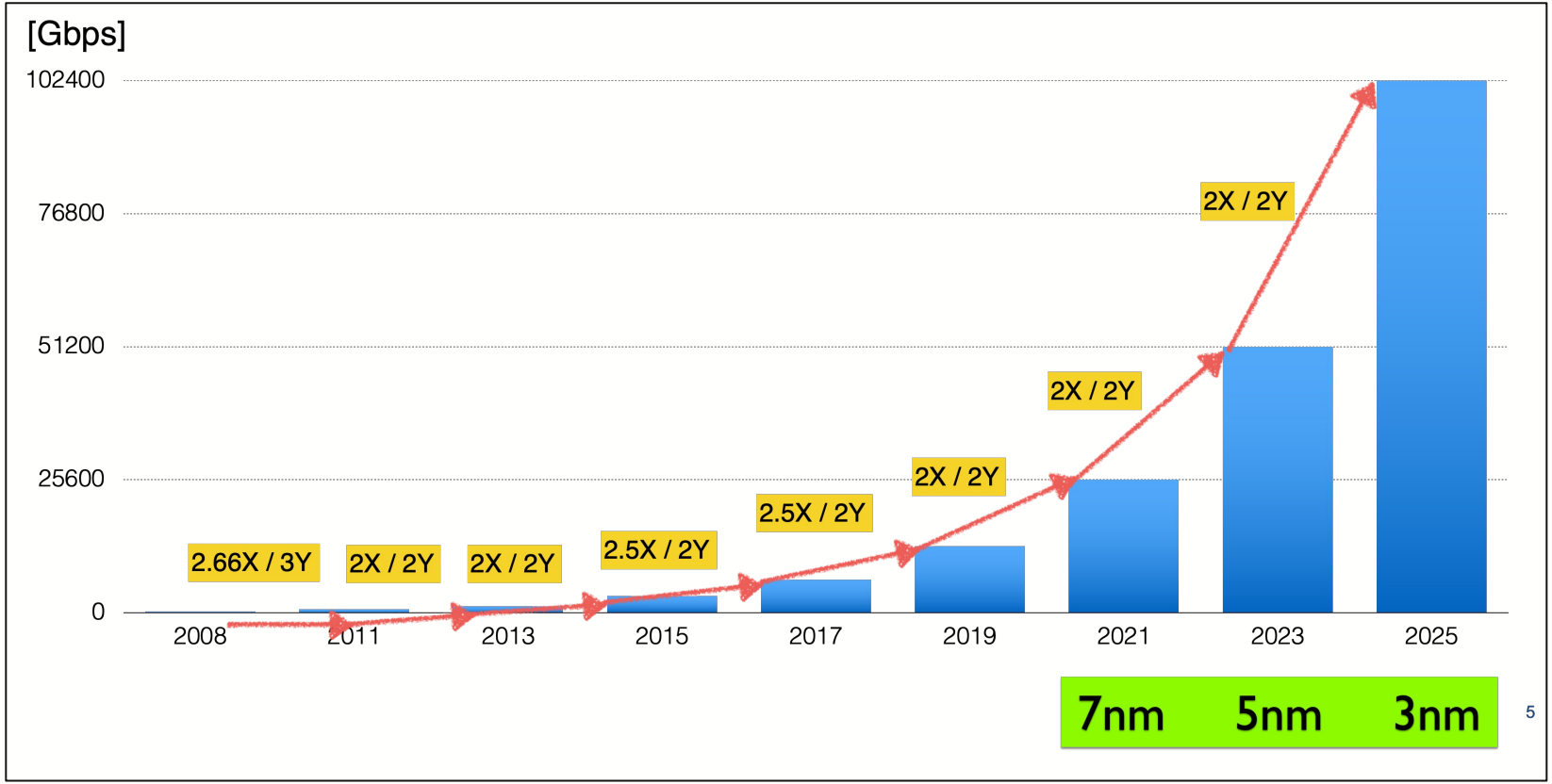
- 半導体技術は進化しており、CPUチップの進化に追従し、スイッチチップも進化していく
- 10年前はかなり遅れていたが、現在は最新のテクノロジーを用いた開発が進んでいる

SoC(System On Chip)の容量変化



- SerDesスピードおよび1チップでサポートできるLane数によりSoC容量が決まる
- ムーアの法則(2年間に2倍)をネットワークに

2025年までの商用スイッチングシングルチップの容量

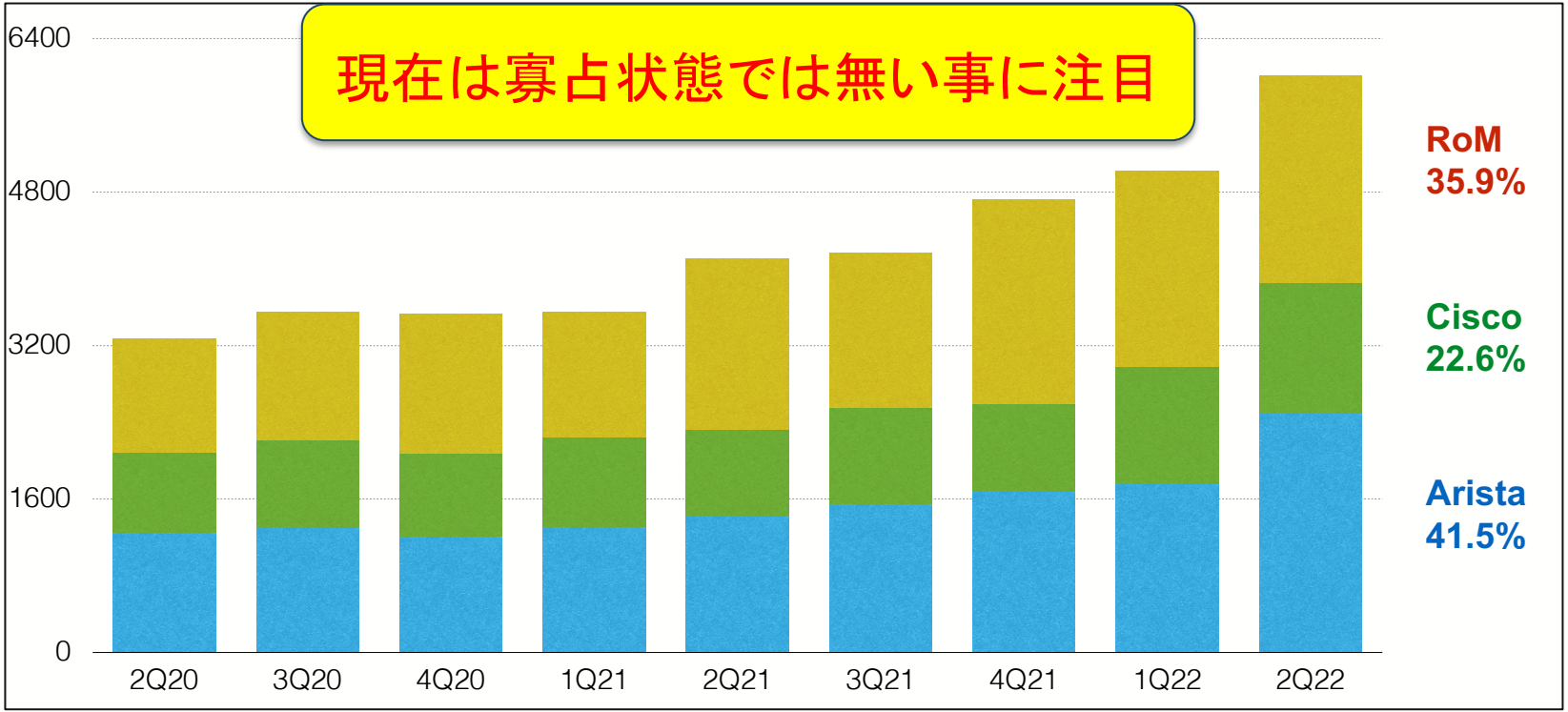


ARISTA

商用シリコンの初のリスト

- 2008年:最初の超低遅延24ポート10Gシングルチップの登場
- 2010年:最初のVoQファブリックを備えたディープバッファ10Gチップの登場
- 2011年:最初の64ポート10Gシングルチップの登場
- 2012年:最初の32ポート40Gシングルチップの登場
- 2013年:最初のVoQファブリックを備えたディープバッファ40Gチップの登場
- 2015年:最初の32ポート100Gシングルチップの登場
- 2016年:最初のVoQファブリックを備えたルータ100Gチップの登場
- 2017年:最初の64ポート100Gシングルチップ登場
- 2018年:最初の32ポート400Gシングルチップ登場(128ポート 100G)
- 2021年:最初の64ポート400Gシングルチップ登場(256ポート 100G)

固定型/モジュラー型スイッチの100G/200G/400Gポートマーケットシェア



Source: Crehan Research Q2'22 Datacenter Switching Report

ARISTA

CMOS Process Nodes through 2025

FinFET
GAA

Process	7nm	5nm	3nm	2nm
TSMC	Now	2021	2023	2025
Intel	Now	2022	2023	2025
Samsung	Now	2021	2024	2025

- TSMCが3nmプロセスを用いた半導体の試作生産を開始、台湾メディア報道
  - https://news.mynavi.jp/techplus/article/20211206-2217419/
- Intelがプロセスの名称を変更、「nm」から脱却へ/10^-10 A（オングストローム）
  - https://eetimes.itmedia.co.jp/ee/articles/2108/03/news066.html
- 5nmプロセス世代のトランジスタが見えてきた「Nanosheet」技術
  - https://pc.watch.impress.co.jp/docs/column/kaigai/1072737.html
- インテルがプロセスの命名規則を変更した理由と今後の展望　インテル CPUロードマップ
  - https://ascii.jp/elem/000/004/064/4064468/

Overall Roadmap Technology Characteristics

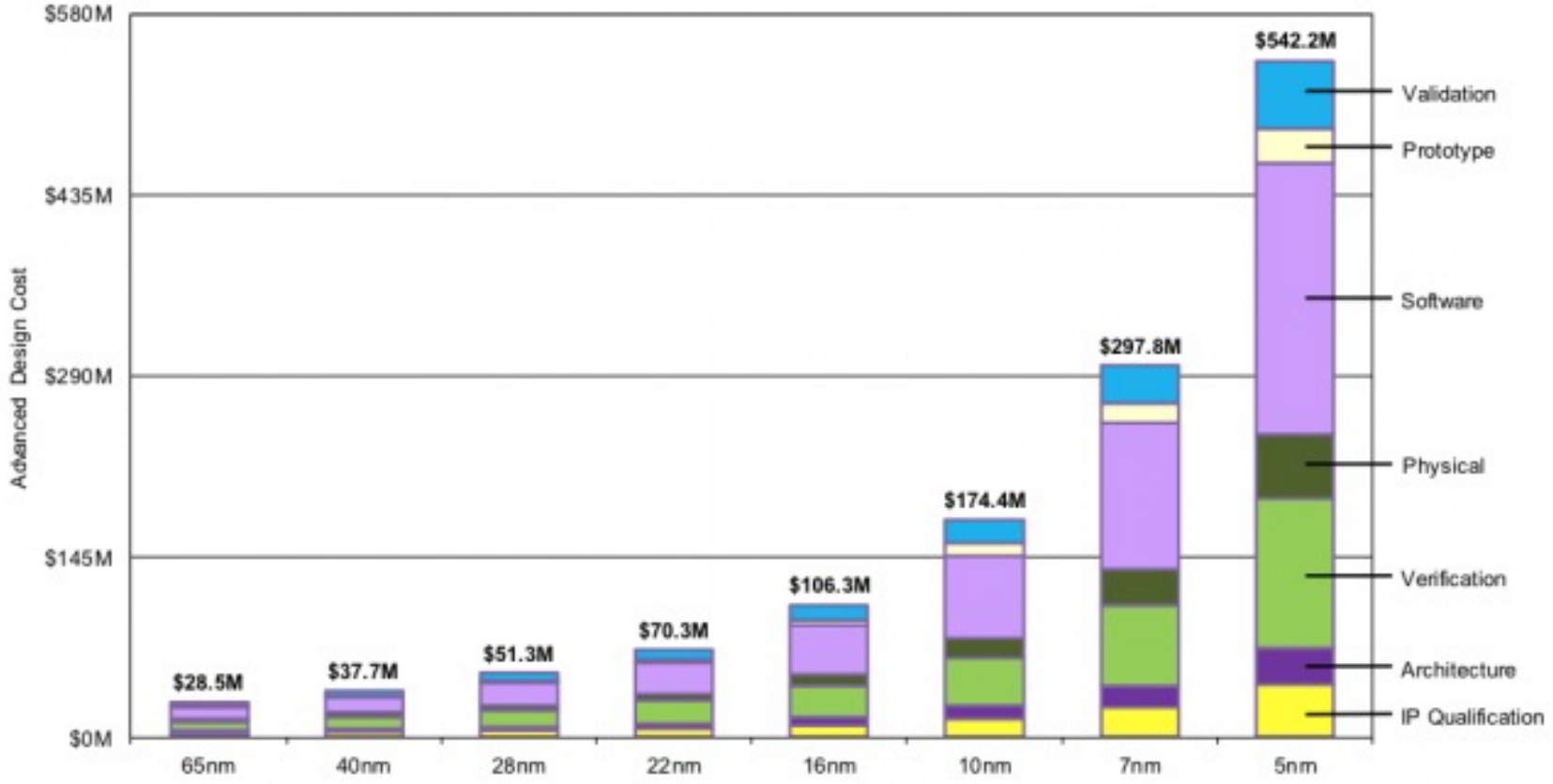
Table ES2 Overall Roadmap Technology Characteristics

2020 IRDS Executive Summary Drivers-ORTC								
YEAR OF PRODUCTION	2019	2020	2022	2025	2028	2031	2032	2034
Logic device technology naming [4] NEW node definition	G54M38	G48M36	G45M24	G45M20	G40M16	G38M16T2	G38M16T3	G38M16T4
Logic industry "Node Range" Labeling (nm)	"7"	"5"	"3"	"2.1"	"1.5"	"1.0nm-eq"	"1.0nm-eq"	"0.7nm-eq"
Logic device structure options	FinFET	FinFET	FinFET LGAA	LGAA	LGAA VGAA	LGAA-3D VGAA	LGAA-3D VGAA	LGAA-3D VGAA
LOGIC CELL AND FUNCTIONAL FABRIC TARGETS								
Average Cell Width Scaling Factor Multiplier	1	0.9	0.9	0.9	0.9	0.9	0.9	0.9
LOGIC DEVICE GROUND RULES								
MPU/SoC M0 1/2 Pitch (nm) [1,2]	18	15	12	10.5	8	8	8	8
Physical Gate Length for HP Logic (nm) [3]	20	18	16	14	12	12	12	12
Lateral GAA (nanosheet) Minimum Thickness (nm)				7	6	5	5	5
Minimum Device Width (FinFET fin, nanosheet, SRAM) or Diameter (nm)	9	7	6	7	6	6	6	6
LOGIC DEVICE Electrical								
Vdd (V)	0.75	0.7	0.7	0.65	0.65	0.6	0.6	0.6
DRAM TECHNOLOGY								
DRAM Min half pitch (nm) [1]	18	17.5	17	14	11	8.4	8.4	7.7
DRAM Min Half Pitch (Calculated Half pitch) (nm) [1]	20.5	17.5	18.5	15	12	10	10	8.5
DRAM Cell Size Factor: aF^2 [4]	6	6	4	4	4	4	4	4
DRAM Gb/1chip target	8	8	16	16	32	32	32	32
NAND Flash								
Flash 2D NAND Flash uncontacted poly 1/2 pitch – F (nm) 2D [1][2]	15	15	15	15	15	15	15	15
Flash Product highest density (independent of 2D or 3D)	512G	1T	1T	1.5T	3T	4T	4T	4T+
Flash Product Maximum bit/cell (2D_3D) [6]	2_4	2_4	2_4	2_4	2_4	2_4	2_4	2_4
Flash 3D NAND Maximum Number of Memory Layers [6]	48-65	64-96	96-128	128-192	256-384	384-512	384-512	512+

THE INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS: 2020

ARISTA

CPUチップの設計コストが高騰中



ARISTA

3nmスイッチングチップの製造コスト

- マスクコスト $25M
- CADツールなど $25M
- 知的財産権 $10M
- 固定コスト $30M
- 人件費 $60M
- トータルコスト $150M

インテルCPUより安いが、それでも高い

経済的実績

1. シリコンビジネスは大量生産でしか成立しない

   先行投資したNRE(Non-Recurring Engineering)コスト
   を回収するためには、10万個以上のチップを売る必要がある

2. スイッチ・シリコンの市場は基本的に安定している

   トップベンダーを駆逐するのは至難の業

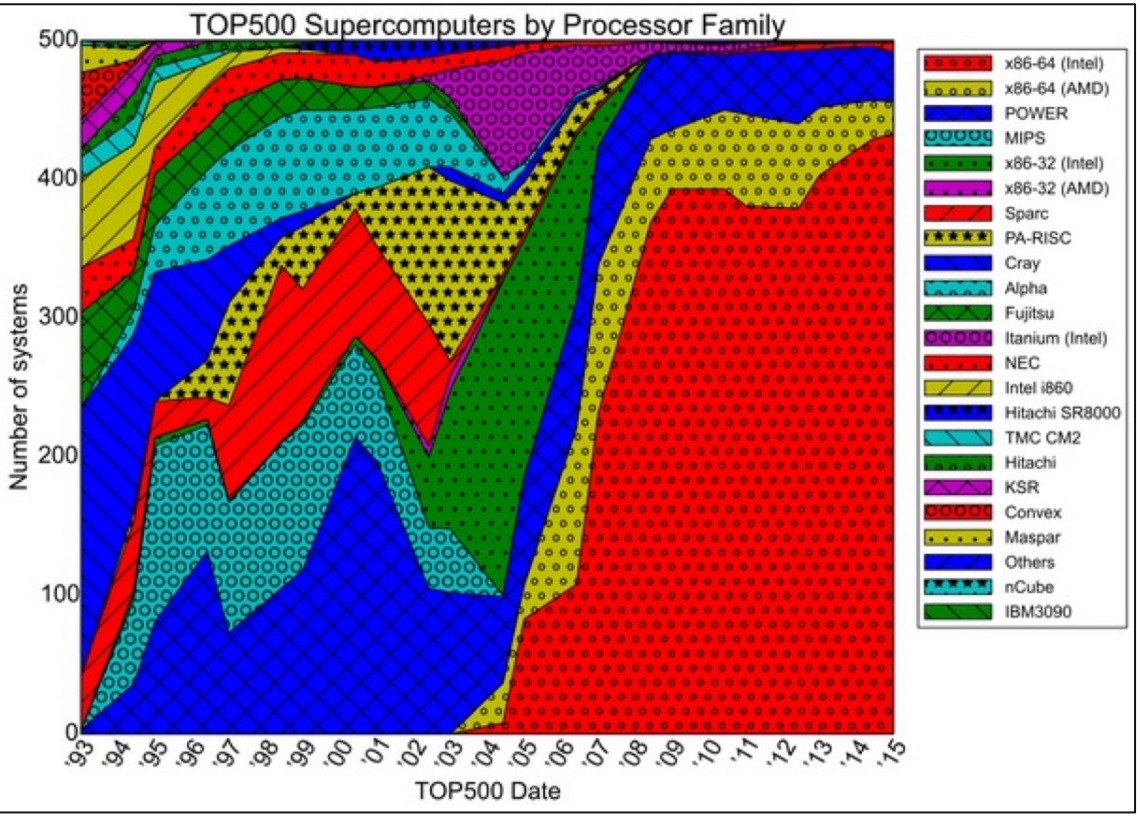
3. 独自チップの製造は経済的ではない

   ボリュームがないと、設計の初期コストが圧倒的に高い

歴史からの類推 1990年代のRISC戦争

各社ともRISCチップのシリコン市場開拓に奔走したが、いずれも失敗に終わった。

なぜ、これらのCPUアーキテクチャは失敗したのか？



大規模投資に見合うマーケットボリュームを確保出来なかった

RISCベンダーは数百万のCPUを販売したがIntelは1億個売った

ARISTA

スイッチ・シリコンの経済性ははるかに厳しい

1990年代のRISC CPUと比較すると、3nmネットワークシリコンの設計コストは桁違いに高く、チップサイズは桁違いに小さい。
チップ数量は1桁少ない
経済性は2桁悪化

ムーアの法則サマリー

1. 3nm以降へのロードマップが明確に
   電力と密度の大幅な改善
2. ネットワーク・スイッチやルーターに大きなメリット
   誰もが低消費電力と高密度のチップを望んでいる
3. ムーアの法則が技術革新を循環させる
   プロセス・ノードが新しくなるごとに、新しいシリコン世代が生まれる
4. シリコンビジネスは大量生産でしか成立しない
   1億5千万ドルの設計コストを正当化するには10万個以上のチップが必要
5. Aristaは高速ポートを誰よりも多く出荷している
   自社でチップを開発することはできない/してない
6. 自社開発のスイッチ・シリコンは高価な趣味
   長期的に見ると、経済性が成り立たない

先端的なシリコンには経済的なボリュームが必要

ムーアの法則がスイッチベンダーにもたらすもの

1. 各世代で密度を2倍に

   2026年までの記録的なプランがロックされる

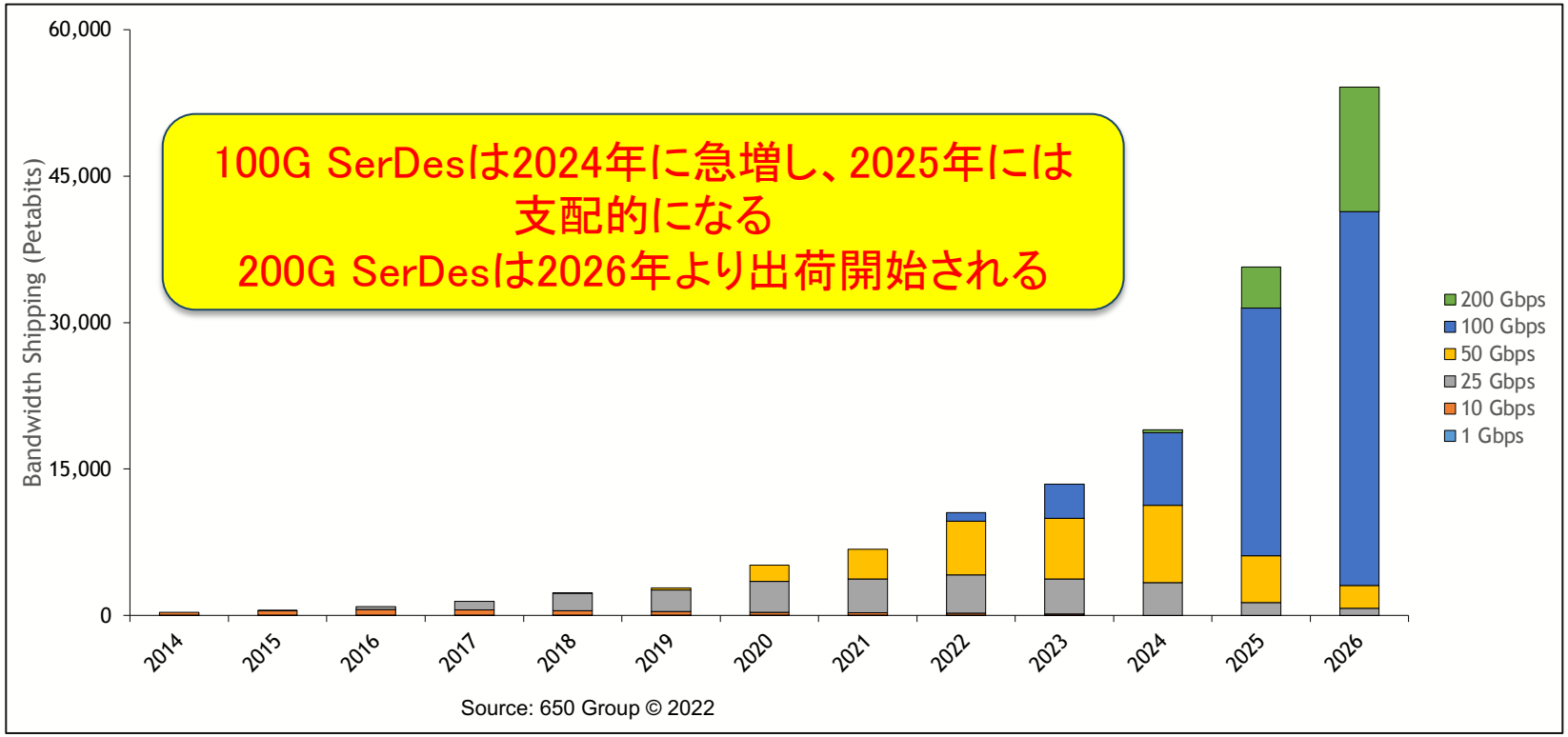
2. プロセスステップを2つ進めるごとに、1ビットあたりの消費電力を半分になる

   7nmから3nmになることで、1ビットあたりの消費電力が半分になる

3. システムレベルの構造コスト低減

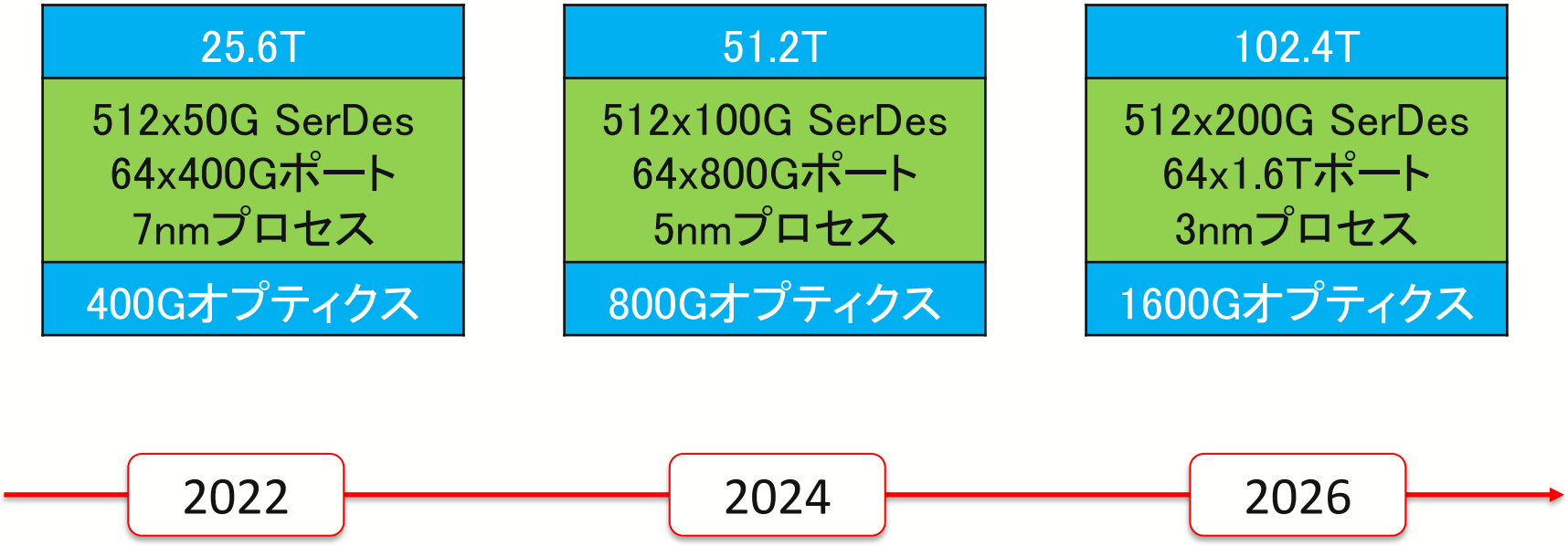
   チップ数、プリント基板数、筐体の小型化

スイッチで出荷されるSerDesのスピード



ARISTA

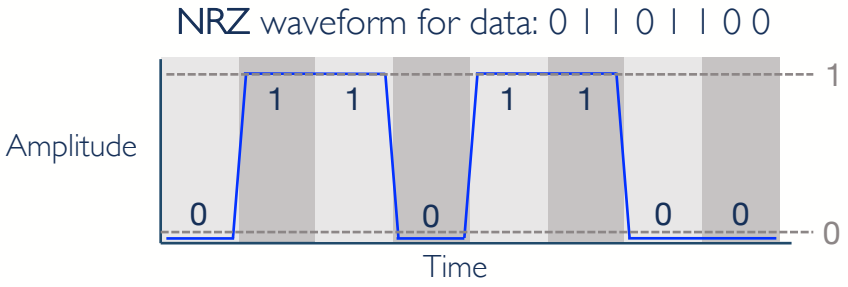
スイッチ用シリコンの密度は2年ごとに倍増

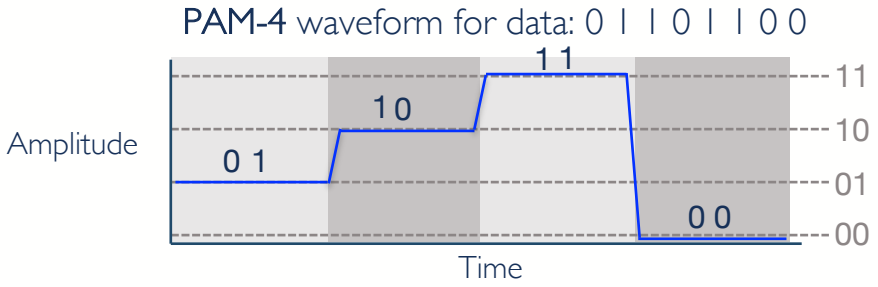


ARISTA

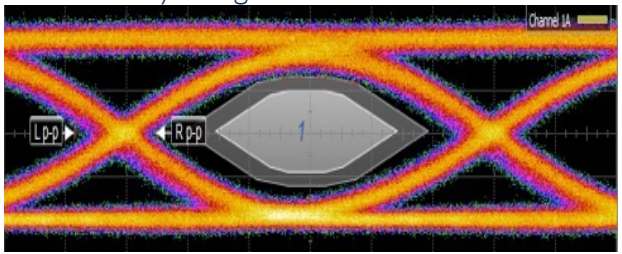
PAM4(Pulse Amplitude Modulation:4パルス振幅変調)

- 400G転送では従来のNRZ(Non Return to Zero)からPAM4に移行
- 01,10,11,00の4つの電圧レベルのパルス信号として伝送

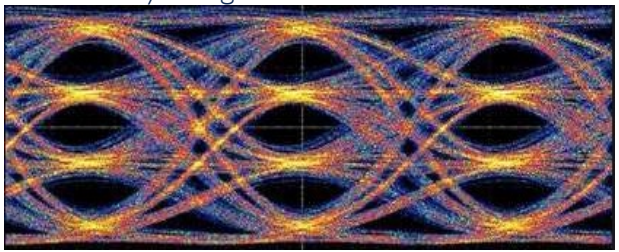




Eye diagram for NRZ data



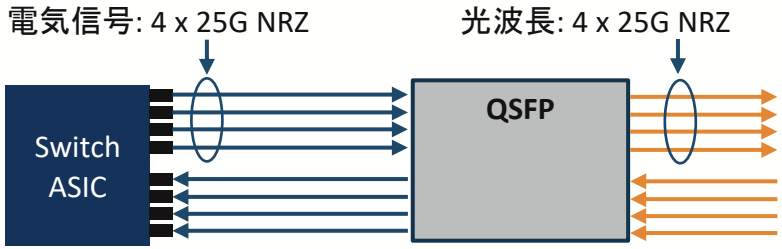
Eye diagram for PAM-4 data



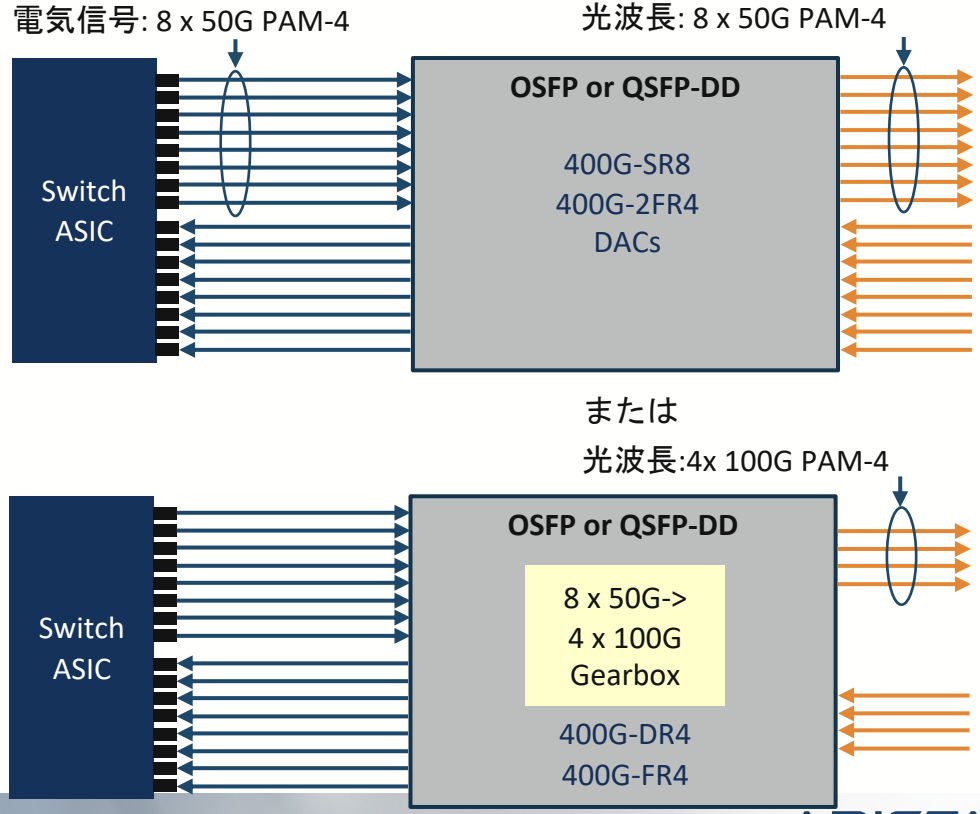
ARISTA

PAM4のインターフェース

既存100Gオプティクス

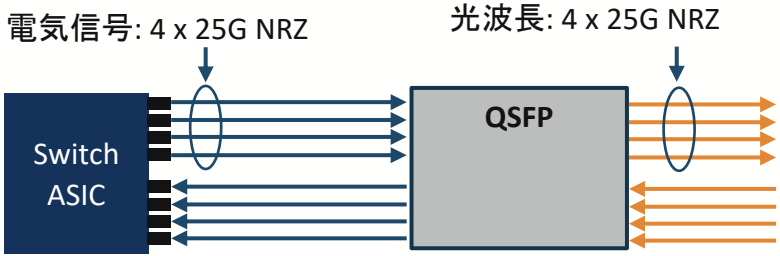


400Gオプティクス

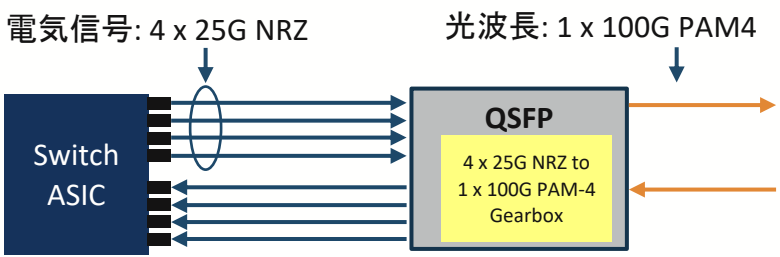


100G DR/100G FRオプティクスモジュール

既存100Gオプティクス

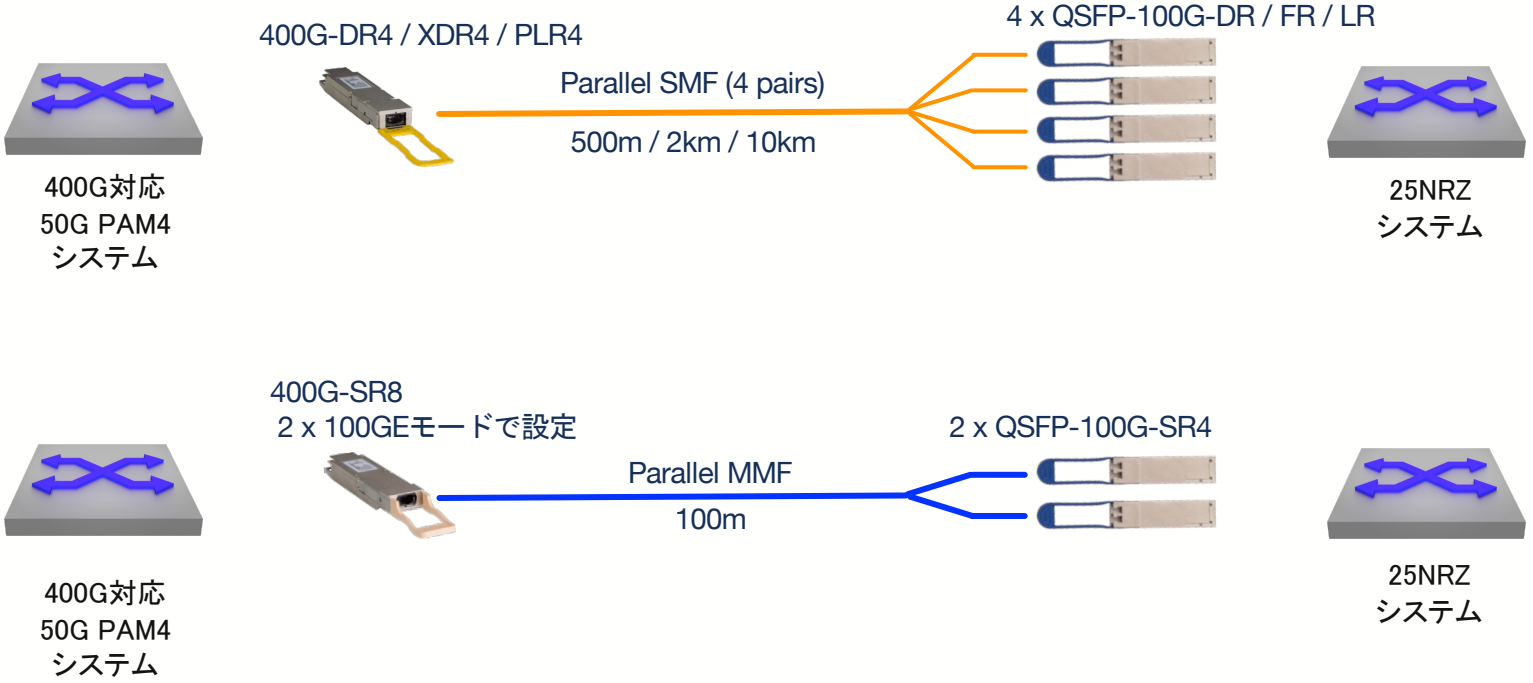


シングルレーン100G-DR/FR/LRオプティクス



- 400Gの最大限活かし、既存の100Gシステムを収容可能な100G DR/FR/LRモジュールがある

400G から 100G QSFPブレークアウトオプション



800Gロードマップ

800Gイーサネット

1. すべての次世代スイッチチップは、800Gイーサネットをサポート

   8x100G SerDesを使用

2. 800Gイーサネットの仕様は2020年に完成

   2020年にイーサネットテクノロジーコンソーシアムによって行われた作業

   https://ethernettechnologyconsortium.org/wp-content/uploads/2020/03/800G-Specification_r1.0.pdf

3. 800Gオプティクスは現在入手可能で、拡大中

   次世代オプティクスがシステムより先に準備できたのは初めて

51.2Tbps 64 x 800G OSFPスイッチ

51.2Tbps
512x100G PAM4
5nm

64ポート 800Gまたは128ポート 400Gを2Uで

- Tomahawk 5 / BCM78900 Series
  - https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm78900-series

なぜ800Gオプティクスか?

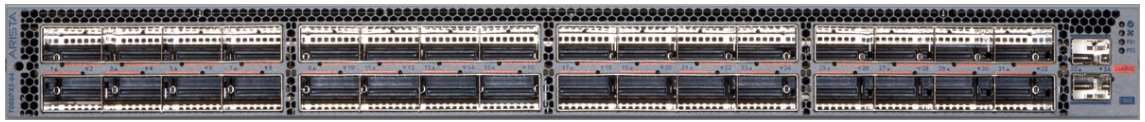
1. 100G SerDesが800Gオプティクスをドライブする
   スイッチのシリコン密度に合わせる必要がある
2. ビット当たりのコスト低減
   オプティクスとシステムレベルでのコスト低減
3. ビット当たりの消費電力低減
   オプティクスとシステムレベルで最大40%の省電力化

一足先に出てきた800Gプラットフォーム

https://www.arista.com/en/products/7060x5-series/specification

25.6T
512x50G PAM4
or 256x100G PAM4
7nm
Tomahawk4

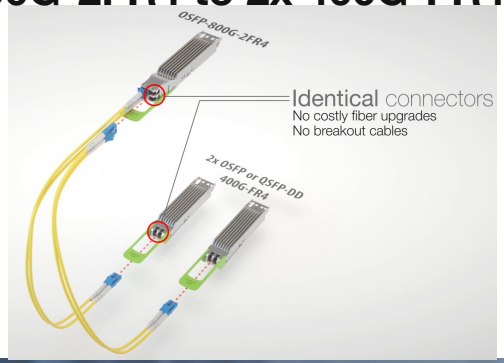
7060DX5-64E
32xQSFP-DD800



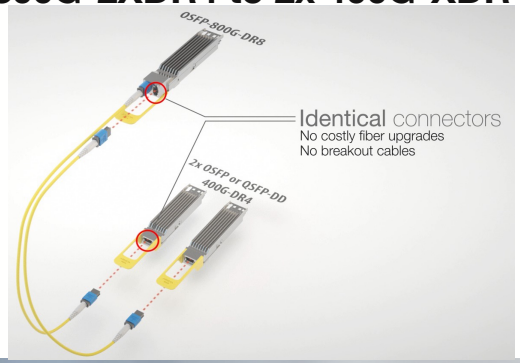
7060PX5-64E
32xOSFP800

- 800GEではなく400GEx2が使われる

800G-2FR4 to 2x 400G-FR4



800G-2XDR4 to 2x 400G-XDR4



ARISTA

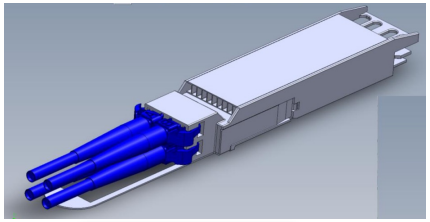
800Gオプティクスモジュール

名称	モジュレーション	距離	コネクタ
800G-ZR	16QAM	100km	LC
800G-FR8/LR8	100G-PAM4	2km/10km	LC
800G-2FR4/2LR4	100G-PAM4	2km/10km	DualLC
800G-DR8/2DR4	100G-PAM4	2km/10km	MPO/2xMPO
800G-SR8/2SR2	100G-PAM4	50m	MPO/2xMPO

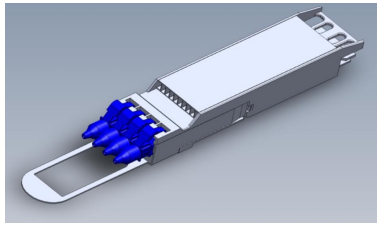
ほとんどの800Gのモジュールはブレークアウトで使用される

800Gブレークアウトアプリケーション

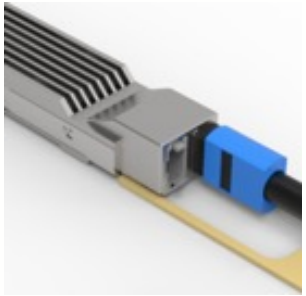
1. 800Gブレイクアウトは多くのユースケースをサポート
   集約、混在、冗長性の向上、より大きな帯域の根本として
2. 800G からDual 400G-DR4/FR4/LR4/ER4 ブレークアウト
   Dual LC、Dual Mini-LCまたはDual MPOでサポート。
3. 800G-DR8から8x100G-DRへのブレークアウト
   Octal SN/MDCコネクターでサポート



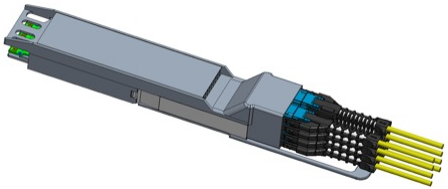
2xLC



2xMiniLC



2xMPO

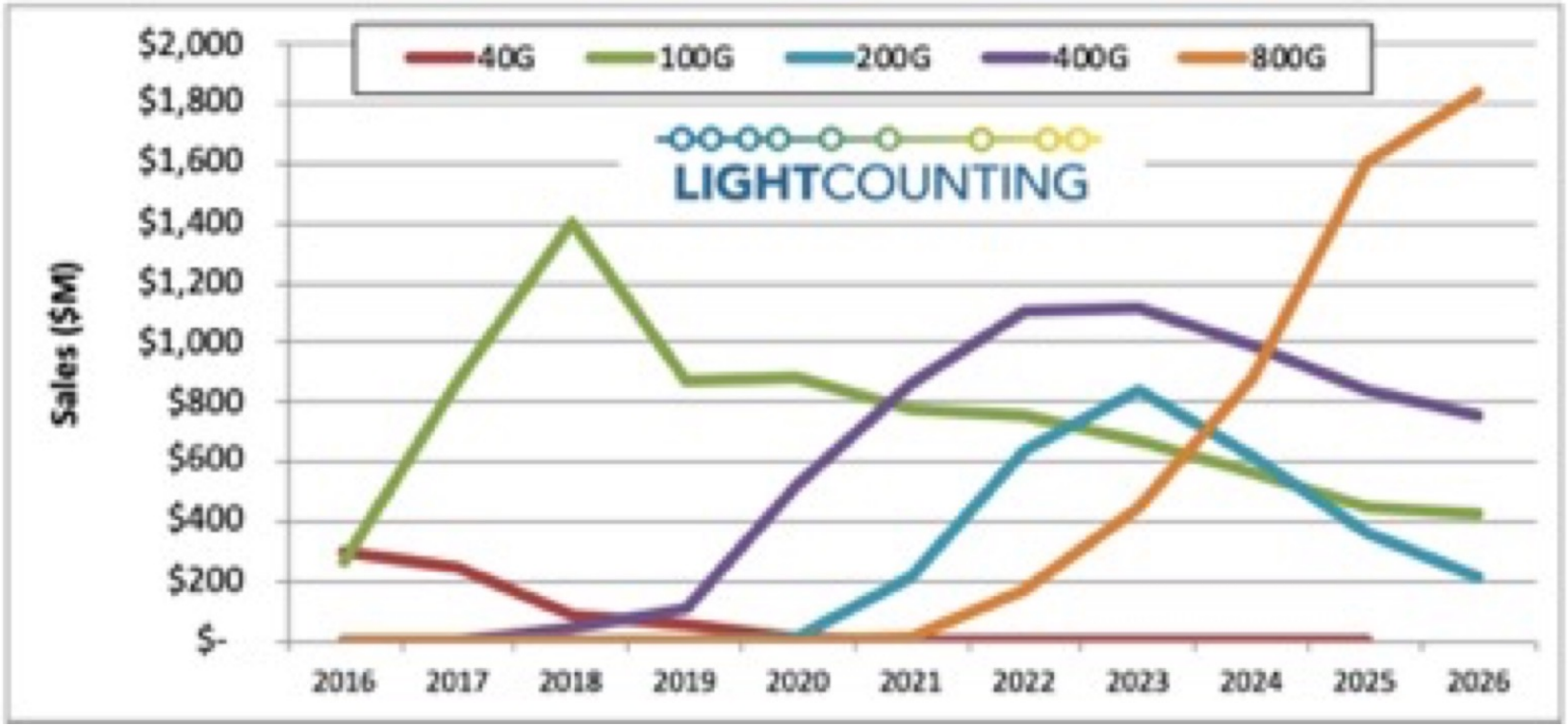


8xSN

800Gデモンストレーション

800Gモジュールの迅速な適用

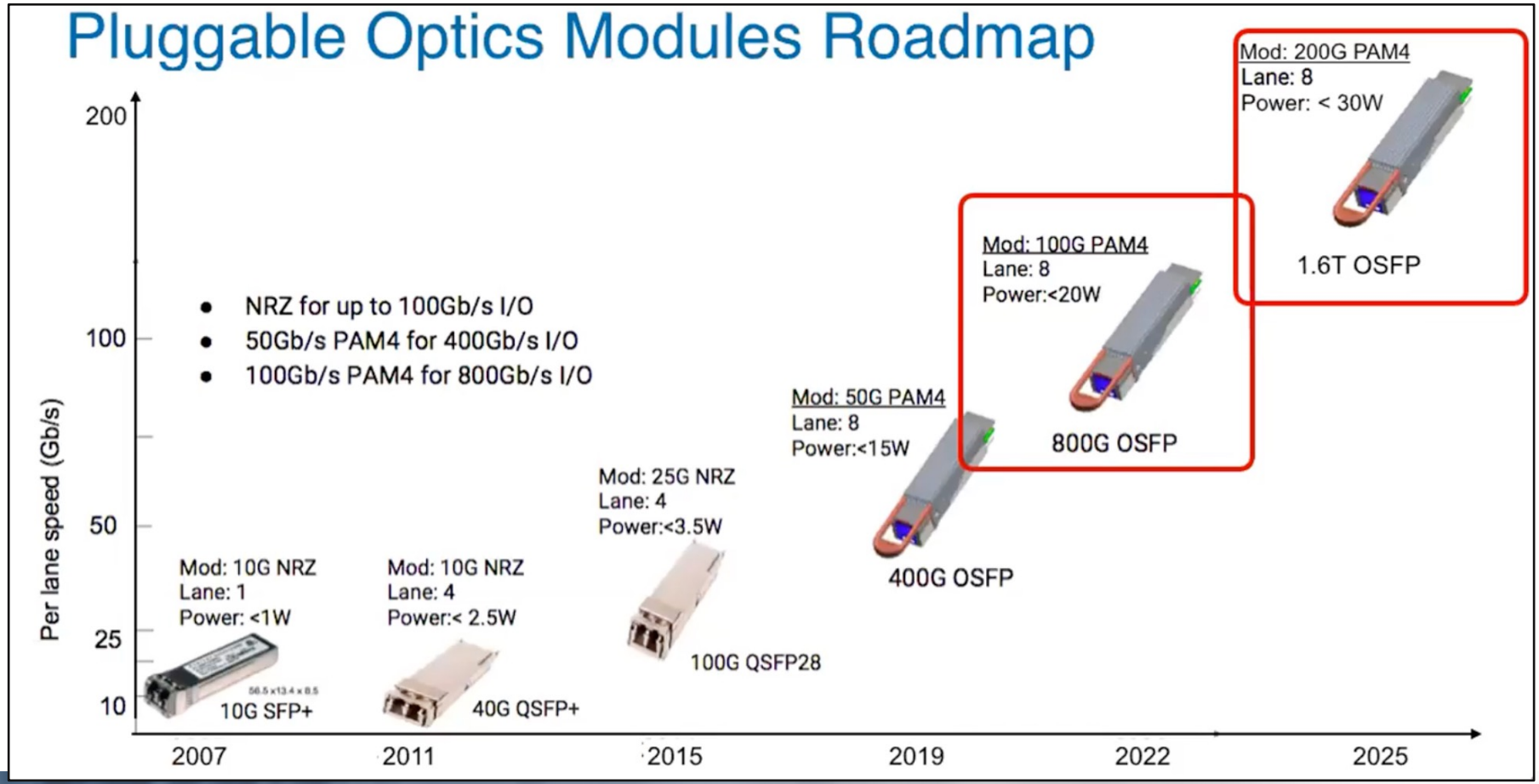
Figure: Sales of Ethernet Transceivers to the Top 5 Cloud Companies



ARISTA

1600Gロードマップ

オプティクスロードマップ



ARISTA

102.4Tスイッチ 64 x 1.6T OSFP 3nm

128x800G または256x400Gを1.6T OSFPでサポート

1.6Tオプティクス(200Gラムダ)のケース

1. ビットあたりのコストと電力を大幅に削減

   正確な数値は未定だが、大幅な削減が可能

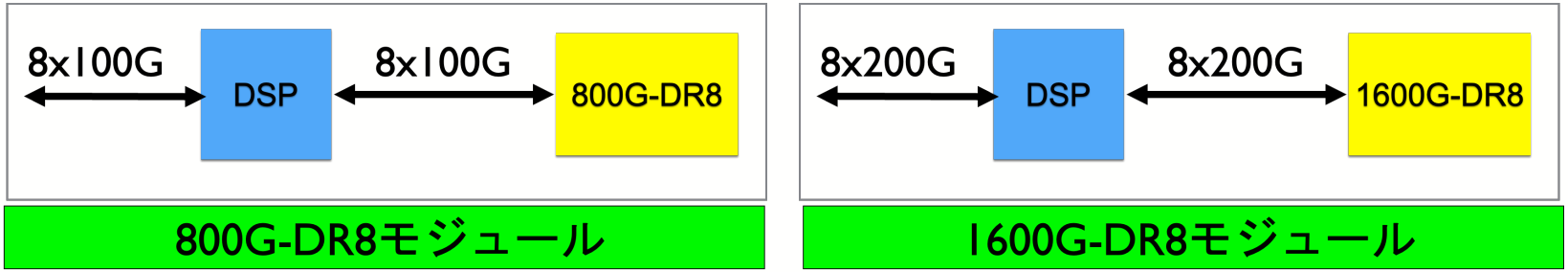
2. フロントパネルあたりの密度を2倍に

   システム構成コストの大幅な削減

3. 将来の200G SerDesを使用した場合の唯一の選択肢

   200G SerDesが1600Gオプティクスモジュールをドライブする

800G-DR8 vs 1600G-DR8



1600G-DR8は800G-DR8同数の部品数
レーザー、変調器、ファイバー・ターミネーション、コネクターは同数

1600Gオプティクスモジュール2026

名称	モジュレーション	距離	コネクタ
1600G-ZR	16-QAM	100km	LC
4x400G-ZR-lite	16-QAM	20km	4xSN
1600G-2LR4	200G-PAM4	6-10km	2xLC
1600G-2FR4	200G-PAM4	1-2km	2xLC
1600G-2DR4	200G-PAM4	2km	2xMPO

ほとんどの1600Gbpsモジュールはブレークアウトで使用される

1600G-OSFPオプティクスモジュール

1. 3nm 200G-PAM4 Serdesエコシステムが牽引する

   5nmチップは全て100G-PAM4となる

2. 主なユースケースは2x800Gと4x400G

   Dual LC、Dual MPO、Quad SNファイバーコネクタを使用

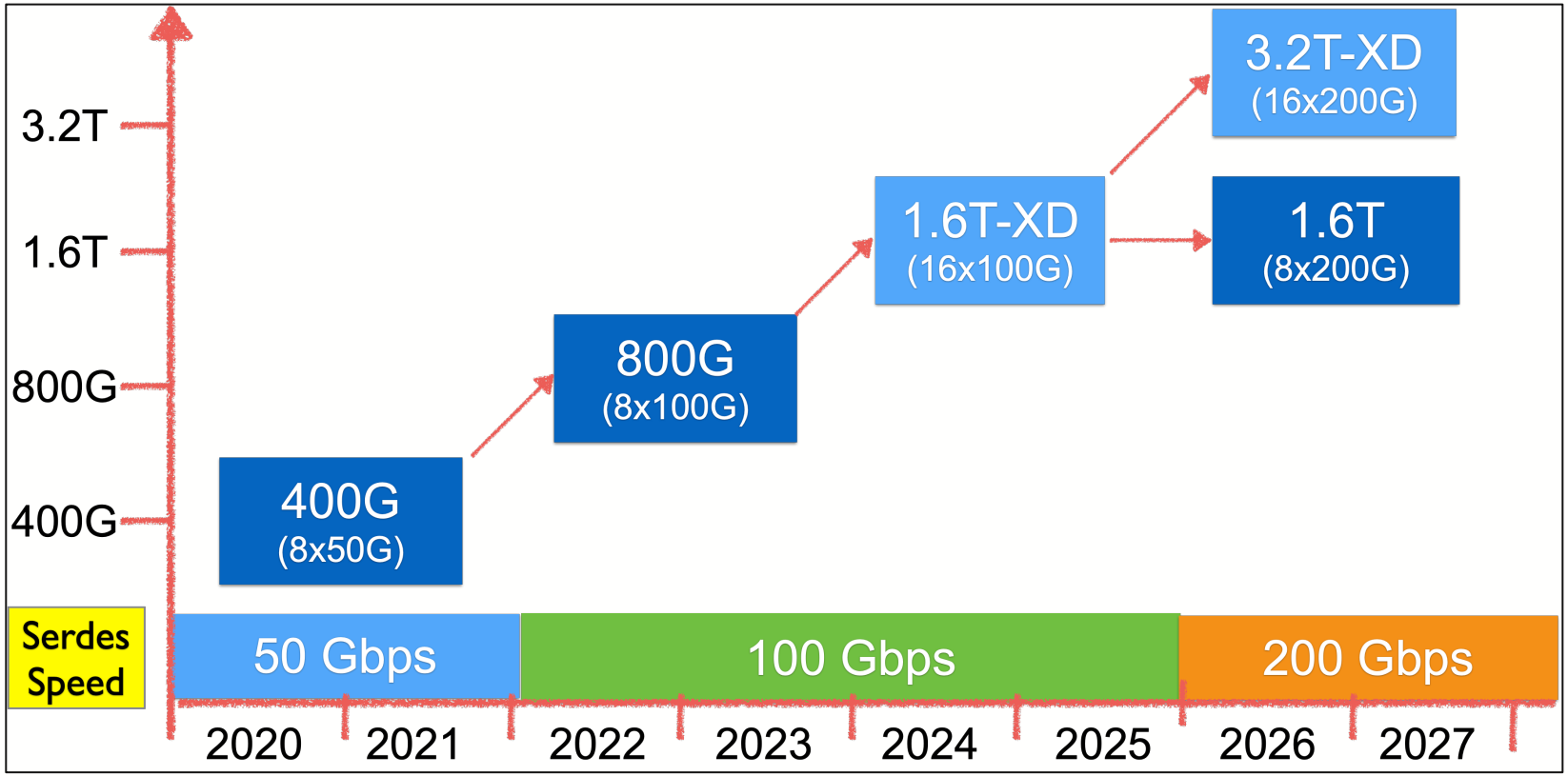
3. OSFP-1600 の仕様が完成

   1600G オプティクスのための堅牢な熱エンベロープ（最大 33W）

1.6Tbpsに早く到達する方法

1. OSFP-XD (eXtra Dense)に入る。
   16Lane 40W電源エンベロープ付き
2. 16x100Gの1600G、16x200Gの3200G
   100G-PAM4エコシステムで1600Gモジュールが使用可能
3. フロントパネルの密度は8レーンに比べ2倍になる
   1U 51.2T（32x1600G）、102.4T（32x3200G）。

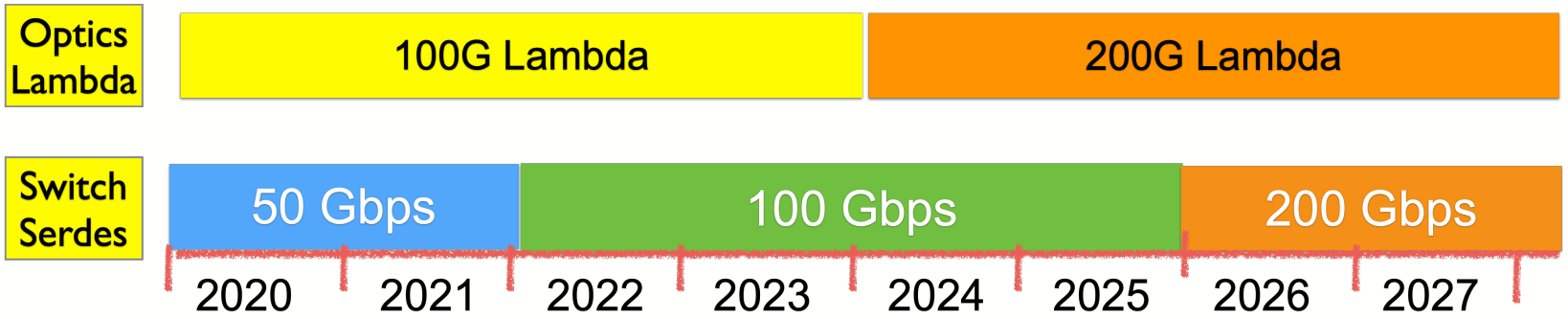
OSFPおよびOSFP-XDロードマップ



ARISTA

電気レーン vs オプティカルレーンスピード

- 100G Lambdaは100G Switch Serdesの2年先を行っていた
- 200G Lambdaは、200G Serdesの2年先を予想



ターミナルエンベロープの重要性

- プラガブルフォームファクタの目的は、光学モジュールに堅牢な熱環境を提供する事
- 熱性能が不十分な場合、故障率が高くなり、消費電力も高くなる

オプティカルモジュールのパワー予想

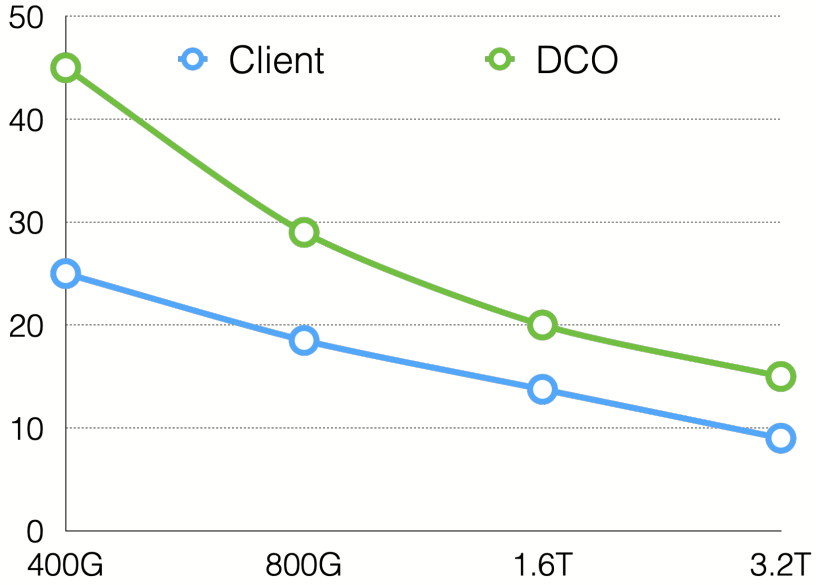
名称	データセンター	DCI –ZR	ZR++
400G(2022)	10-12W	16-20W	20-24W
800G(2024)	15-18W	22-24W	30W
1600G(2026)	20-24W	30-36W	40W+
3200G(2028)	35-40W	TBD	TBD

注）消費電力は導入年度の値

プラグマブルオプティクスの電力進化

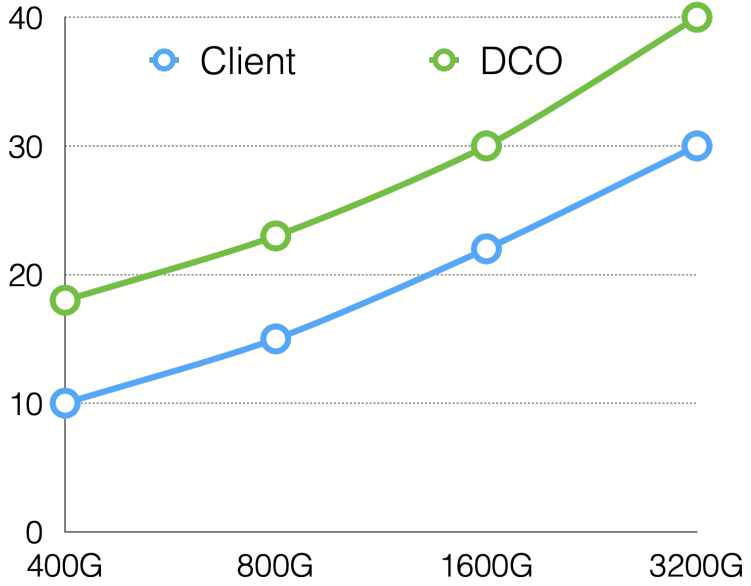
1bit伝送する為に必要なエネルギー

pJ/Bit



モジュール当たりに必要な電力

Watt per Module



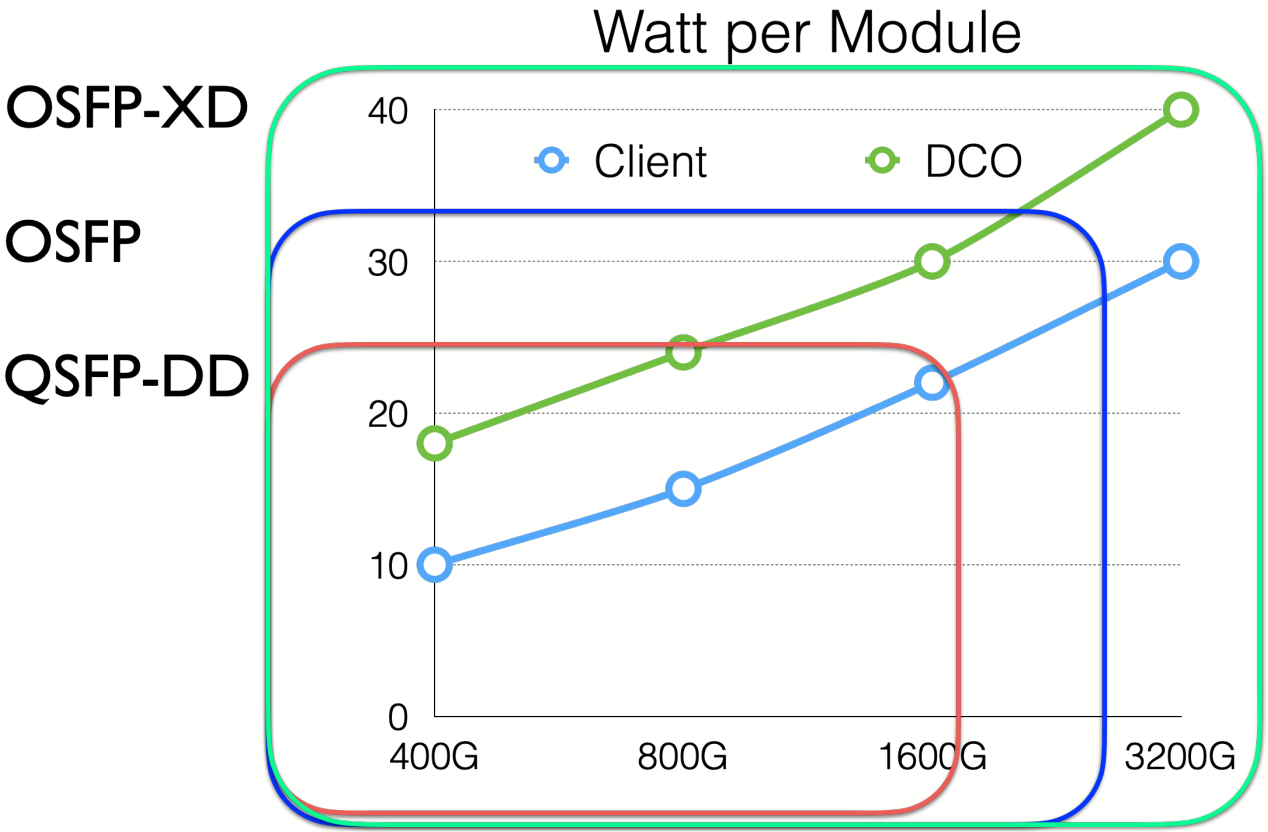
プラガブルオプティクス パワーエンベロープ

QSFP-DD
400G(8x50G) 800G(8x100G)
25Wまで

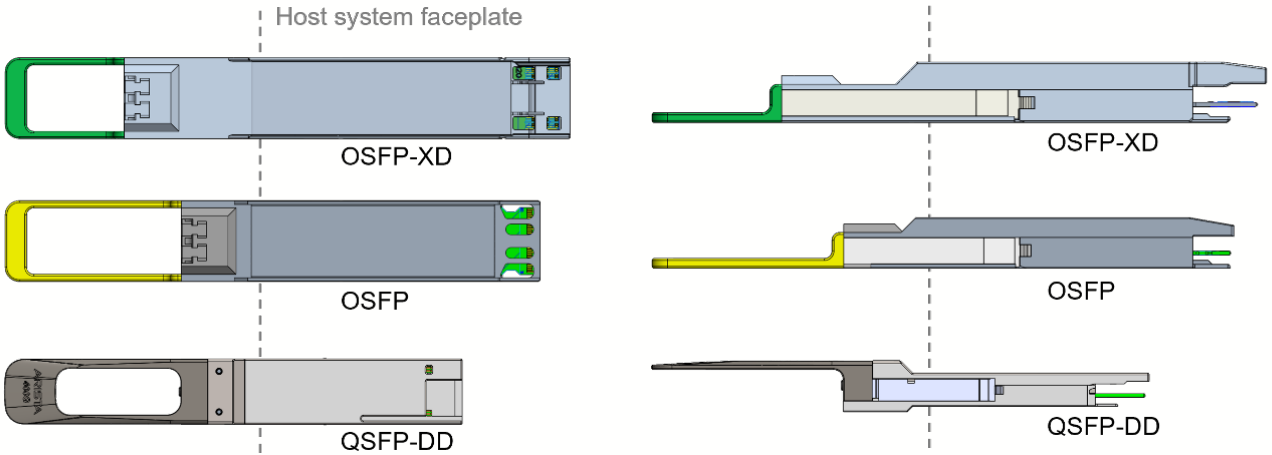
OSFP
400G(8x50G) 800G(8x100G) 1600G(8x200G)
33Wまで

OSFP-XD
1600G(16x100G) 3200G(16x200G)
40Wまで

プラガブルオプティクス熱外装

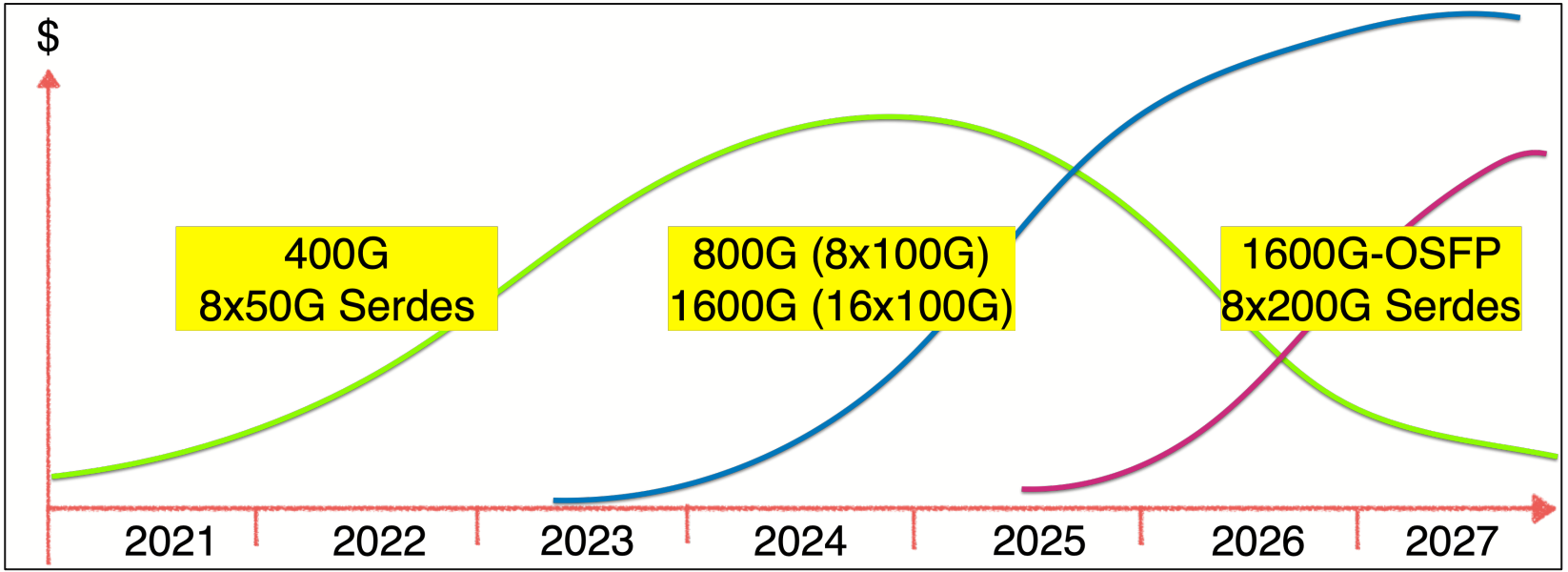


光学モジュールのフォームファクターのサイズ比較



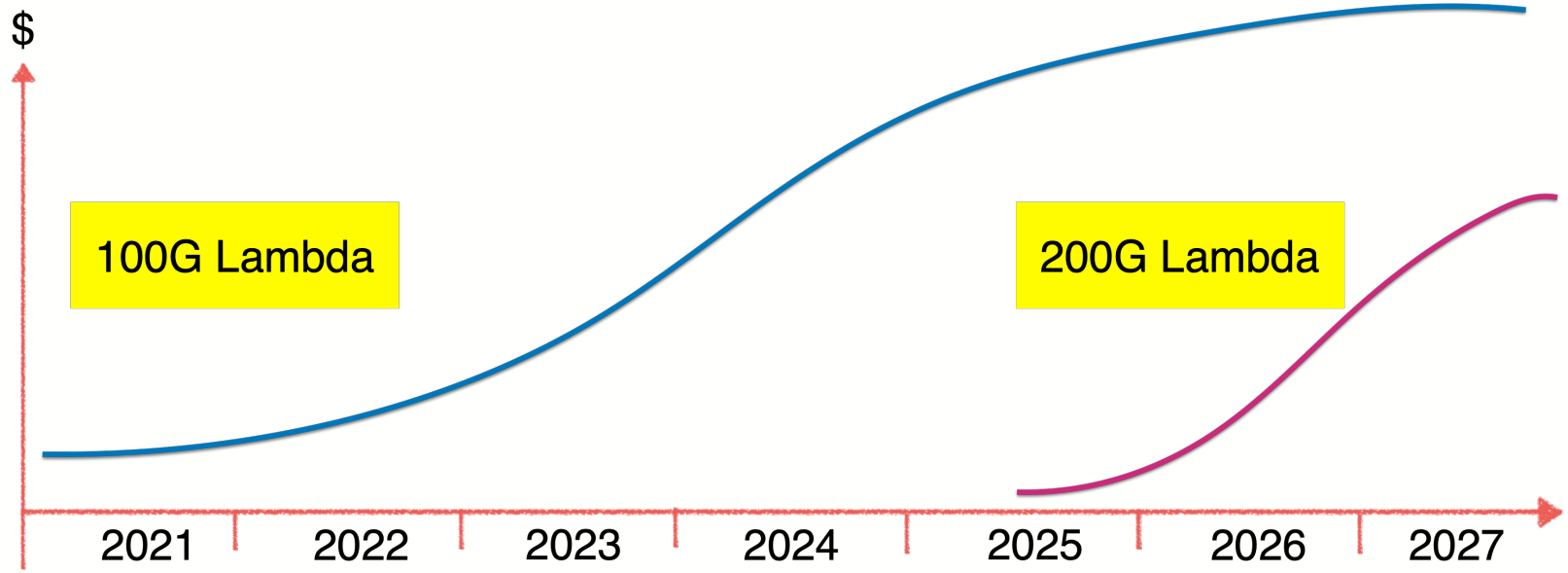
内蔵されたヒートシンク+広い表面面積=高熱伝導率化

400G-800G-1600フォームファクター移行予想



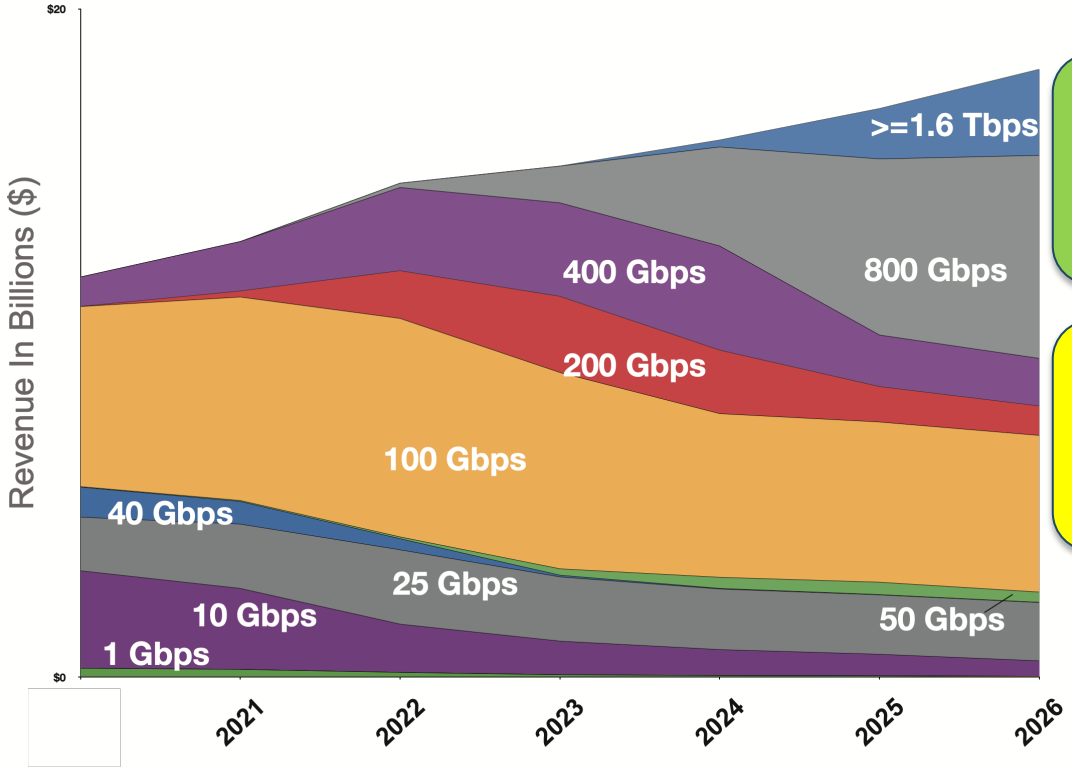
スイッチシリコンの高密度化により、
フォームファクタの高密度化が進む

100G λ から200G λ への移行



200Gラムダの採用は2026年に急増する

400Gから800G、1600Gへの移行予測



2026年には800Gと1600Gが市場全体の45%を占める見込み

2026年には400G以下が全体の55%を占めると予想

Source: Dell'Oro March 2022 - Long Term Ethernet Switch Forecast

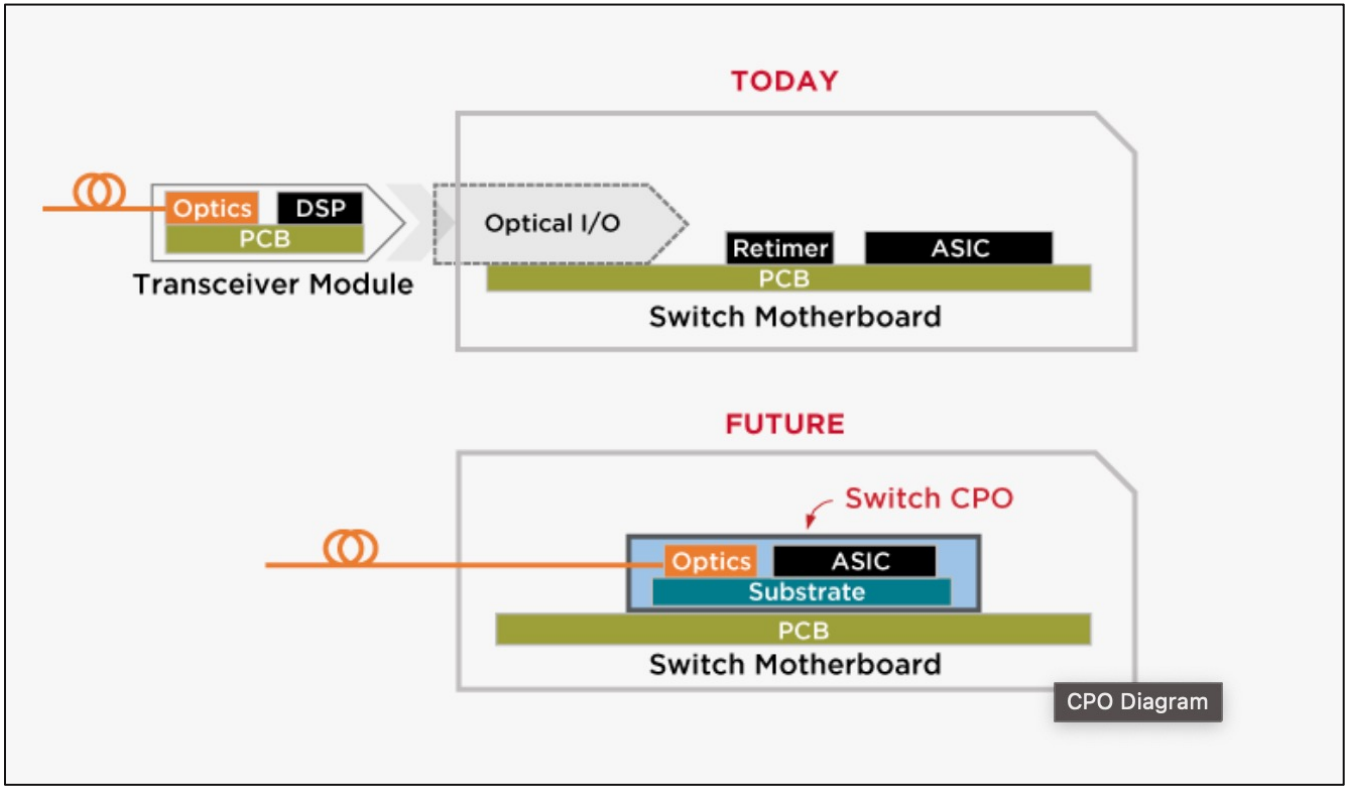
ARISTA

サマリー

1. 800G、1600Gポートへの急速な移行
   ビット当たりのコストと消費電力の削減が原動力
2. 800Gおよび1600Gのオプティクスがもたらす大きな利点
   ビットあたりのコストと電力を削減
3. ほとんどの800G/1600Gオプティクスはブレイクアウト・モードを使用
   Dual 400G/800GおよびOctal 100G/200Gポート
4. OSFP と OSFP-XD が 1.6T への唯一の現実的な道(現時点で)
   QSFP-DDは熱外囲が十分でない。
5. 200G λ オプティックスには2025年以降に開始
   100Gから200G λ への移行は数年後
6. 51.2T や 102.4T では、Co-Packaged Opticsは必要ない。
   システムの大部分はプラガブルオプティクスを使用する。

Co-Packaged Optics (CPO)

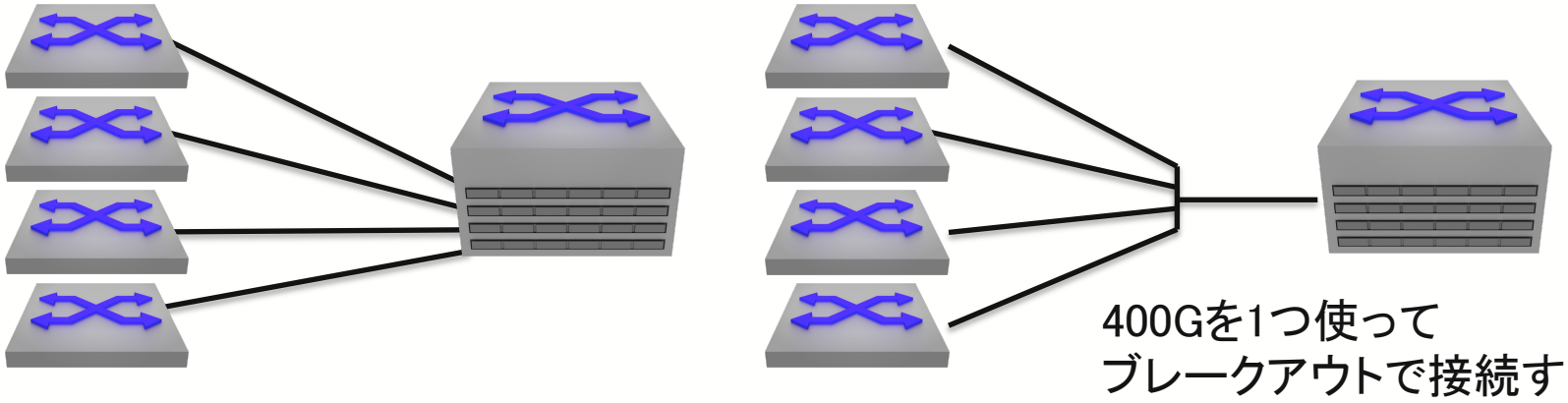
https://www.broadcom.com/info/optics/cpo



ここで

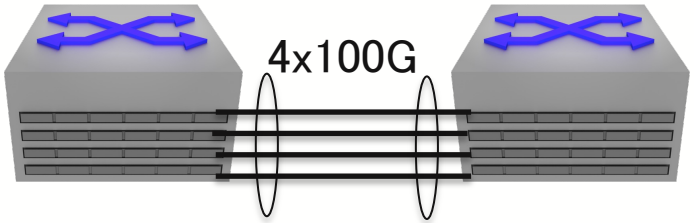
ブレークアウトとか使わねーよ！！とか思ってますか？

ブレークアウトの活用/100G->400Gを例に

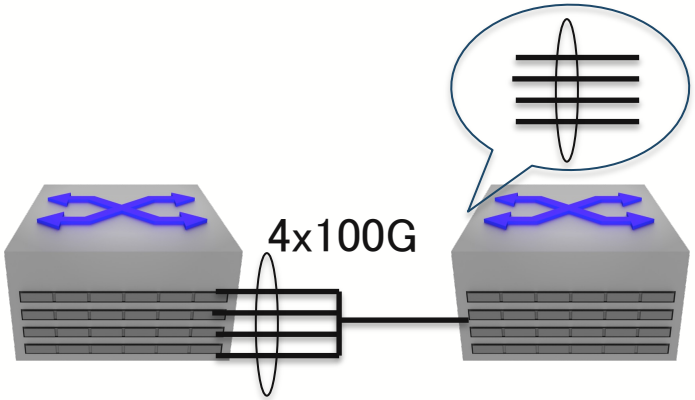


- インターフェースの速度が変わるとシンプルに使用するポート数が少なくなる
- ブレークアウトを使うと収容出来る対抗数を増やす事が出来る

ブレークアウトの活用/100G->400Gを例に



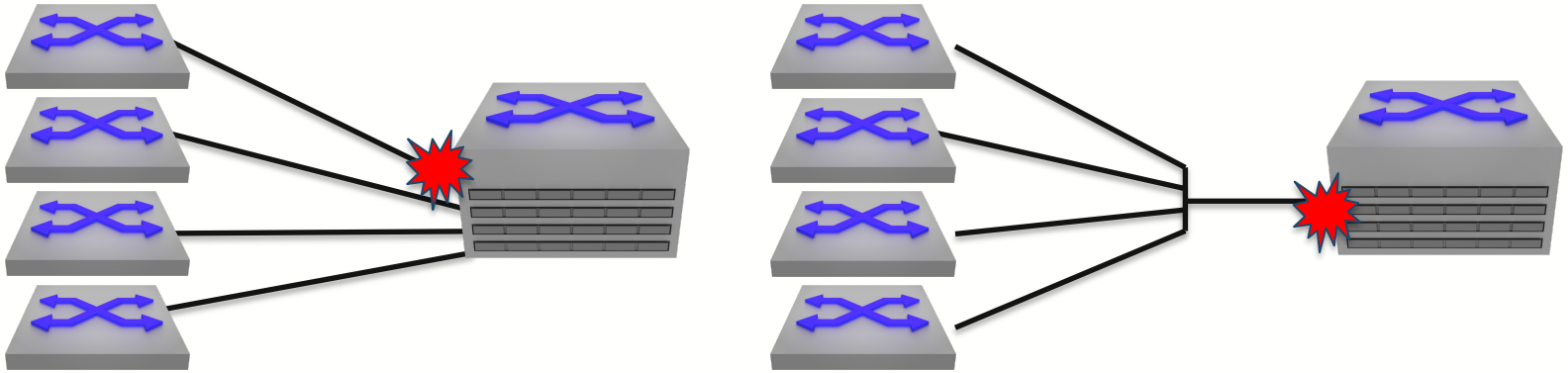
双方4ポート使用



400Gを1つ使って
ブレークアウトで接続する

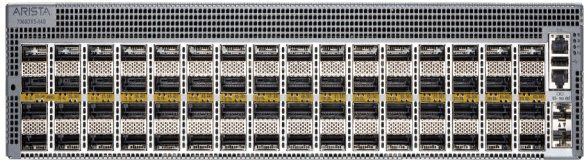
- 特にインターネットトラフィックなどインターフェースでは帯域が足りない。冗長性と増速のためにポートチャネルをする
- ポート数をへらす事も出来る(4x100G->400G)
- 対抗の機器はそのままで1つのポートでポートチャネルをする事も出来る

ブレークアウトの活用/障害時



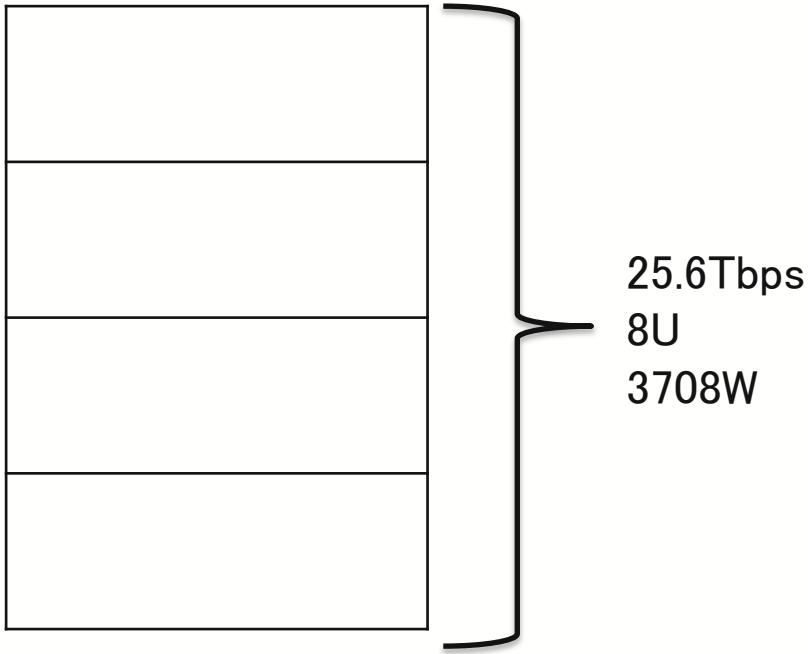
- 100Gのトランシーバーが故障する場合1対抗のみの障害になる
- ブレークアウトの400Gが故障すると4対抗全部障害になる
- 信頼性の高いもの使いたくなるけどな。。。。

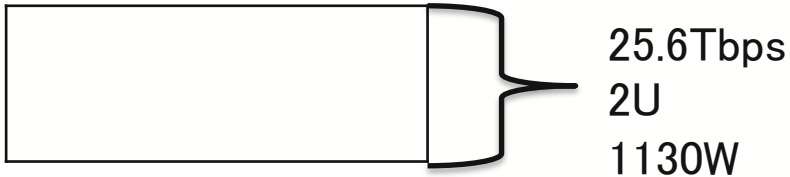
消費電力比較



	7260CX3-64	7060DX5-64S	-
IO	64x100G	64x400G	4倍
帯域	6.4Tbps	25.6Tbps	4倍
ラックサイズ	2U	2U	同じ
チップセット	Tomahawk-2/16nm	Tomahawk-4/7nm	-
消費電力	314 W / 927 W	985 W / 1130 W	1.2倍
100Gサポート数	64	256	4倍
100G当たりの消費電力	14.5W	4.41W	1/3
リリース	2017	2022	5年

ラックサイズ比較





- 過去の機器での帯域がこれだけのスペースや電力で使用出来るようになる

議論したい事

- ブレークアウトを利用するメリットに比べて他にデメリットってありますか？
- どのような状況で新しいインターフェース検討しますか？
  - 帯域の使用率
  - 価格の安定
  - 実績
  - 外部からの圧

参考資料

- The Road to 800G and Beyond Arista Andreas Bechtolsheim
  - https://youtu.be/krEEnHjfvRQ
- Arista 800G Transceivers and Cables: Q&A
  - https://www.arista.com/assets/data/pdf/Arista-800G-Transceivers-and-Cables-FAQ.pdf
- Arista 400G Transceivers and Cables: Q&A
  - https://www.arista.com/assets/data/pdf/Datasheets/Arista-400G_Optics_FAQ.pdf
- The Next Generation of Pluggable Optical Module Solutions from the OSFP MSA
  - https://osfpmsa.org/assets/pdf/OSFP1600_and_OSFP-XD.pdf