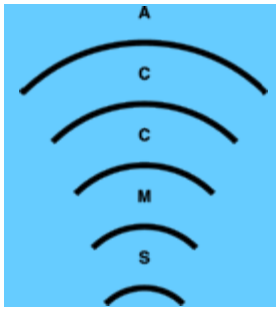


Exa-Scale Computingへの道

中島　浩

（京都大学／HPCIコンソーシアム理事）

目次

- 道の遠さ&険しさ
  - Exa-Scale の見通し@2010
  - 遠さ&険しさの度合い: 性能x100@...
- 道程を見通す活動@日本
  - 全体像
  - それぞれの役割
- 現時点での見通し@日本
  - 体制像
  - Post 京システム像 (技術的)
  - Post 京システム像 (理念的)

道の遠さ&険しさ
Exa-Scaleの見積@2010

year		2009 Jaguar	2011 京	2018	2018/ 2011
sys. peak perf.	[PF/s]	2	11	1000	O(100)
sys. power	[MW]	6	15	20	O(1)
sys. memory	[PB]	0.3	1.4	50	O(10)+
node perf.	[TF/s]	0.13	0.13	1-10	O(10-100)
node mem b/w	[GB/s]	25	64	~1000	O(10)
node conc.		48	64	10^3-10^4	O(10-100)
node inj. b/w	[GB/s]	3.5	20	200-400	O(10)
total #nodes	[x10^3]	19	88	100-1000	O(1-10)
total conc.	[x10^6]	0.9	5.6	1000	O(100)
storage	[PB]	15	10+	~1000	O(10)+
storage b/w	[TB/s]	0.2	0.7	60	O(100)
sys. MTTI	[day]	~10	~10?	~1	O(1/10)

based on presentation by P. Beckman in IESP WS @ Oxford, 2010
http://www.exascale.org/mediawiki/images/7/75/IESP-Oxford-Intro-Beckman.pdf

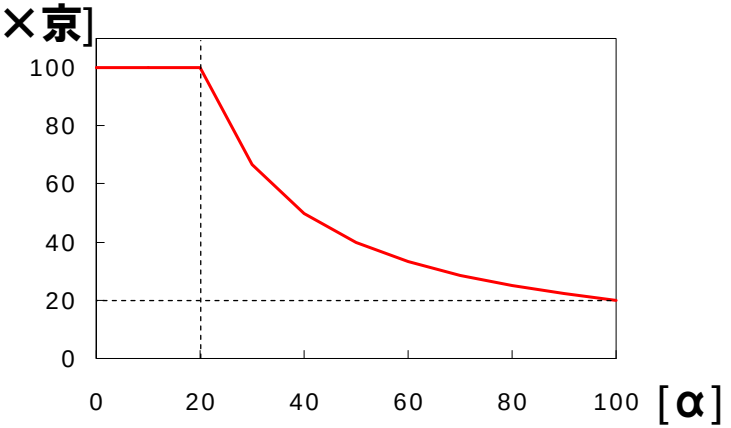
遠さ＆険しさの度合い (1/3)

- 作る険しさ: **性能**x100@**電力**x1
  - 性能／電力比
    - SPARC64 VIIIfx (**京**): 128GF/s@60W = 2.2GF/s/W
    - PowerPC A2 (BG/Q): 205GF/s@55W = 3.7GF/s/W
    - Sandy Bridge: 166GF/s@115W = 1.4GF/s/W
    - Xeon Phi: 1TF/s@225W = 4.5GF/s/W
    - Exa-Target: 1EF/s@20MW ➔ 50GF/s/W ✕ 2?
  - Xeon Phi vs Exa-Target node

		Xeon Phi	Exa-Target	ratio
perf.	[TF/s]	1000	1-10	O(1-10)
mem b/w	[GB/s]	320	~1000	O(1)+
conc.		960	10^3-10^4	O(1-10)
inj. b/w	[GB/s]	15	200-400	O(10)+
power eff.	[GF/s/W]	4.5	~100	O(10)+

遠さ＆険しさの度合い (2/3)

- **使う険しさ: 演算性能x100@メモリ性能x10+**
  - **メモリ／演算性能比**
    - SPARC64 VIIIfx (京): 128GF/s@64GB/s = 0.5B/F
    - PowerPC A2 (BG/Q): 205GF/s@43GB/s = 0.2B/F
    - Sandy Bridge: 166GF/s@51GB/s = 0.3B/F
    - Xeon Phi: 1TF/s@320GB/s = 0.3B/F
    - Exa-Target: 10TF/s@1TB/s ➔ 0.1B/F
  - **仮に 0.1B/F とすると**
    - **メモリバンド幅使用率=α%@京のプログラムの(理想)性能**
      (実質α≒100%が多数）



遠さ＆険しさの度合い (3/3)

- 作る＆使う険しさ: 性能x100@並列度x100
  - 並列度x100≒トランジスタ数x100

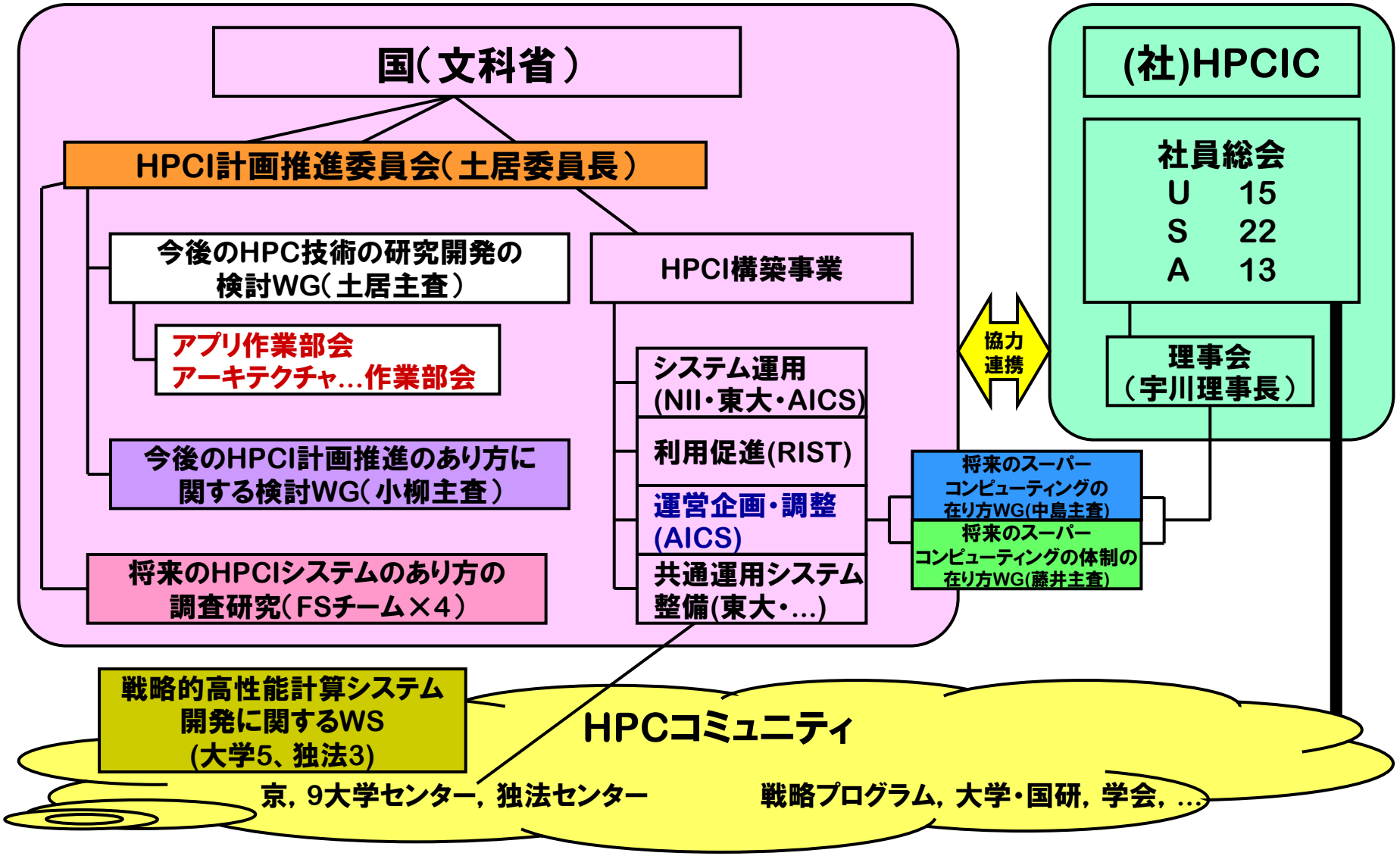
    ➔故障率x1 という訳にはいかない (たとえばx10)
    - 故障率x100にしない工夫
    - 故障率x10 (MTTI<1日) でも運用する工夫
    - 故障率x10 でもアプリが長時間走る工夫
  - 並列度x100の源泉は?
    - weak scaling (問題サイズx100) ➔ 実行時間＞(or≫)x1
    - strong scaling ➔ 遅延影響↑, 通信／計算比↑,
      non-trivial な並列性利用
    - wide SIMD ≒ vector but ≠vector

      ← cache に当てなければ話にならない

      連続データでなければ話にならない

      vector の苦手 (short vec., 条件節, ...) は共有

道程を見通す活動@日本
全体像



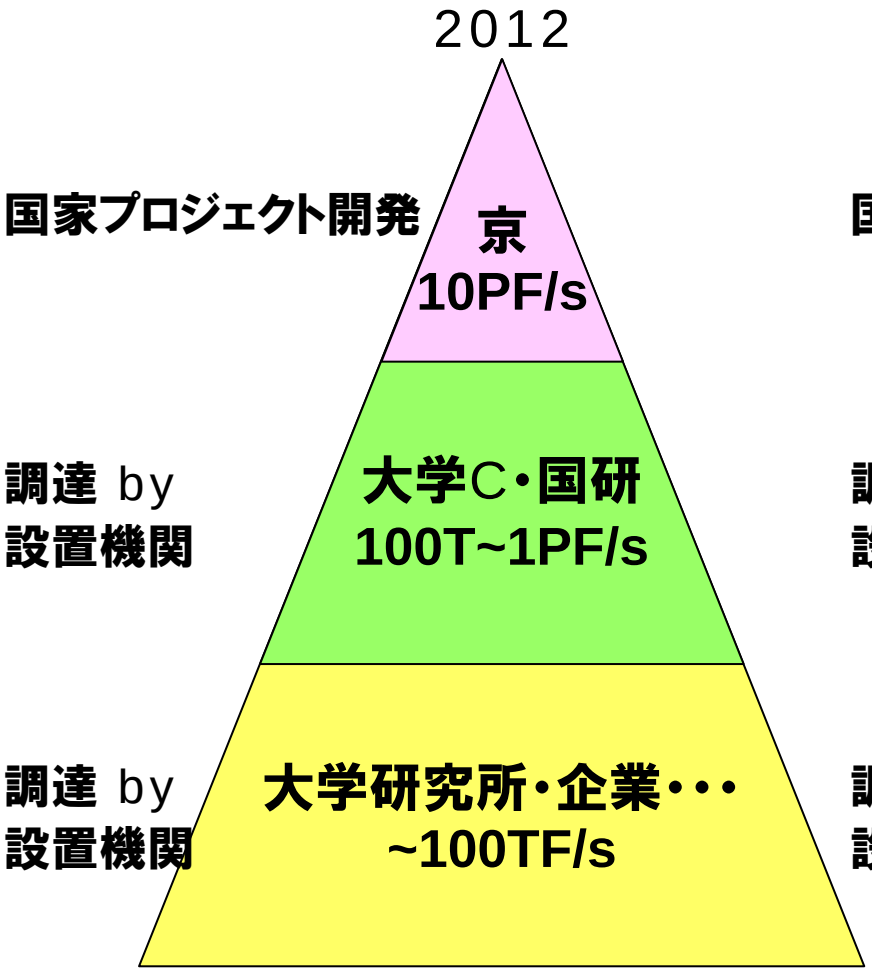
それぞれの役割 (1/2)

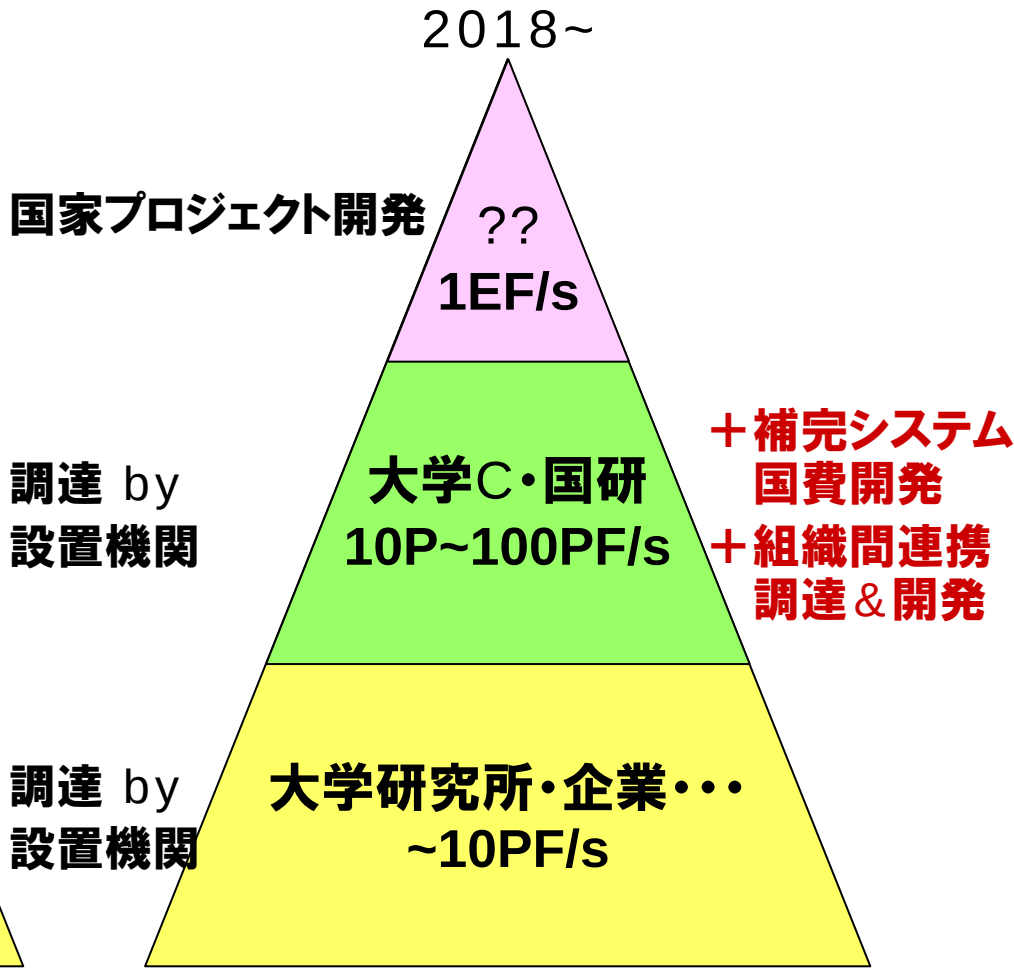
- HPCI計画推進委員会(土居委員長+9委員, '10~)
  - 文科省HPC施策の取りまとめ
- 技術も含む全般的検討
  - 文科省
    - 今後のHPC技術の研究開発の検討WG (土居主査+18委員, '11): 現状＆課題分析, 検討課題整理
    - 今後のHPCI計画推進のあり方に関する検討WG (小柳主査+24委員, '12~13): 論点整理, システム＆体制像
  - 理研＋HPCIコンソーシアム
    - 将来のスーパーコンピューティングの在り方に関する検討WG (中島主査+7委員, '12~13): post 京システム像 (理念的)
    - 将来のスーパーコンピューティングの体制の在り方に関する検討WG (藤井主査+6委員, '12~13): HPCI体制像

それぞれの役割 (2/2)

- 技術検討
  - 戦略的高性能計算システム開発に関するWS ('10~)
  - 土居WG作業部会 ('11)
    - アプリケーション作業部会: アプリ@Exa, 要求性能パラメータ
    - コンピュータアーキテクチャ・コンパイラ・システムソフトウェア作業部会: post 京候補システム像 (技術的), 技術ロードマップ
  - 将来のHPCIシステムのあり方の調査研究 ('12~13)
    - アプリケーション分野から見た … (代表:富田@理研AICS)
      - 科学的成果@Exa, ベンチマーク by 代表的Exaアプリ
    - レイテンシコアの高度化・高効率化による … (代表:石川@東大)
    - 演算加速機構を持つ … (代表:佐藤@筑波大)
    - 高バンド幅アプリケーションに適した … (代表:小林@東北大)
      - 各システムの設計パラメータ検討, アプリ性能予測

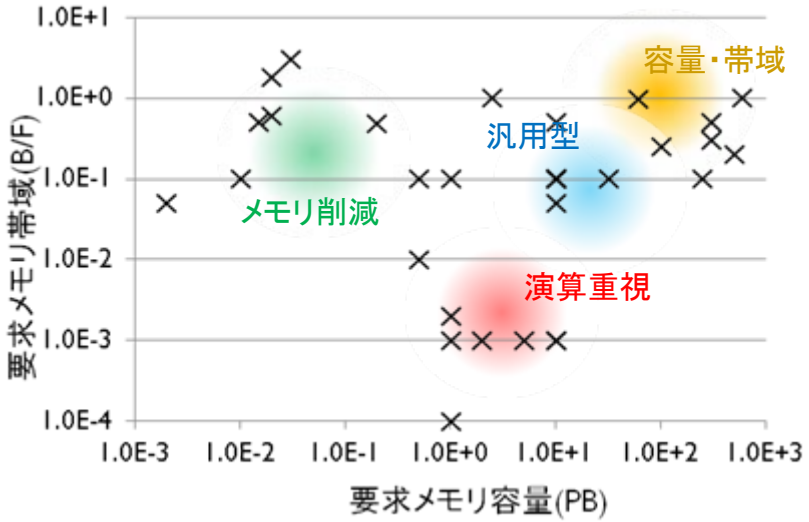
現時点での見通し@日本
体制像

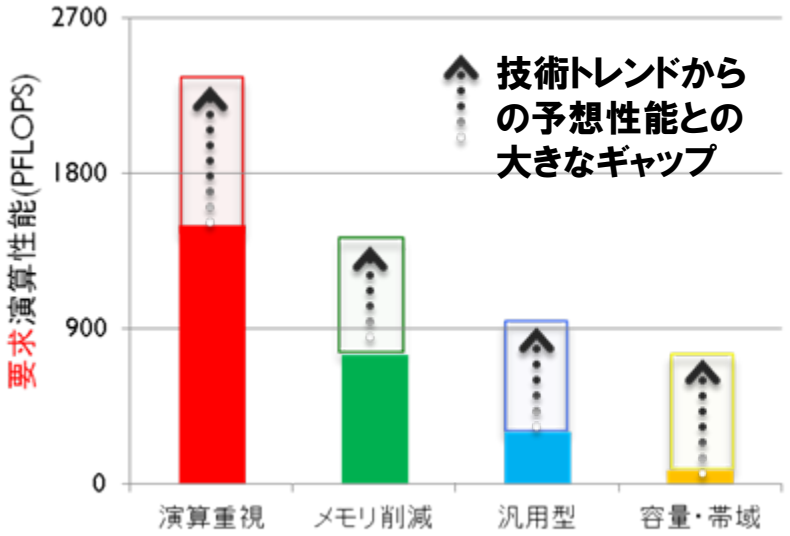




現時点での見通し@日本
Post 京システム像 (技術的)

	peak perf. [PF/s]	mem. b/w [PB/s]	mem. cap. [PB]	inj. b/w [GB/s]
IESP	1000	~100	50	200-400
汎用型	200- 400	20- 40	20 - 40	30-150
容量帯域重視	50- 100	50-100	50 -100	
メモリ容量削減	500-1000	250-500	0.1- 0.2	
演算重視	1000-2000	5- 10	5 - 10	





source: **今後のHPCI技術開発に関する報告書(概要)**
http://www.mext.go.jp/b_menu/shingi/chousa/shinkou/028/shiryo/1321887.htm

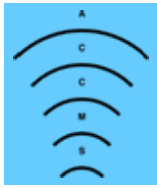
現時点での見通し@日本

Post 京システム像 (理念的・1/2)

- 開発の必要性・意義
  - 科学技術成果を創出する最先端研究開発装置
  - 先駆的ハード・ソフト技術を具現化するプラットフォーム
    - 日本で自立的・継続的に保有・開発すべき技術
    - 国際的標準利用・標準形成のための源泉技術
- システム完成時期・規模・性能・アーキテクチャ
  - 完成時期＝2017~2018 (≒京の引退次期)
  - peak perf. target＝1EF/s (w/ 科学的成果裏付)
  - アーキテクチャ: 汎用型＆単一 vs 適合型＆複数
    - 複数開発困難➔単一が必然➔広範囲のアプリ適合が必要
    - 多様なアプリ対応 by 第2階層, 適合型メカニズム部分導入, …
  - アーキテクチャ＆多様アプリ対応策の判断基準
    - アプリ性能／電力, 初期成果・実現コスト, 技術継続性・波及効果, 世界的HPC技術トレンド, 下方展開, 後続開発・整備計画

Post 京システム像 (理念的・2/2)

- **技術開発要素**
  - **ハードウェア** (**プロセッサ**, **メモリ階層**, **結合網**, I/O, …)
    - **選択・集中** based on **技術的優位性**, **ハードコスト**, **ソフトコスト**, **国産技術推進**, **将来的発展性**, **技術投資回収**, …
    - **国産プロセッサ** w/ competitive cost/performance
  - **ハード＋システムソフト＋アプリソフト三位一体課題**
    - B/F**低下**➔**システムソフトでの補償＋B抑制アルゴリズム開発**
    - **故障率増加**＆**性能電力比相対的低下**
      ➔**ハード＋OS協調**, **システムソフト技術** for **アプリ対応**
  - **システムソフト** (mainly for **アプリプログラミング**)
    - **コンパイラ技術➔アーキテクチャ進化・変遷の影響最小化**
    - **ライブラリ＋フレームワーク➔アーキテクチャ進化・変遷への耐性**
    - **アルゴリズムレベルでの並列度拡張サポート**



(とりあえずの) まとめ

- Exa-Scale への道
  - かなり険しい due to 収穫逓減, no free lunch, ...
  - 先送りしても険しさは緩和されない (だろう)
  - あきらめるとそれっきりになる (可能性大)
- 登攀挑戦@日本
  - 1EF/s@2018 に向けて動き出しつつある (ような感じ:-)
  - あらゆる stake holder (incl. yourself) にとっての commitment chance
- 険路登攀法 for app. people
  - がむしゃらに攀じ登る
  - いずれできるリフトの予約をする
    - ライブラリ／フレームワークを使う・作らせる・作る予算を取る・・・
  - 魔法の絨毯をひたすら待つ