Beyond Latency and Throughput

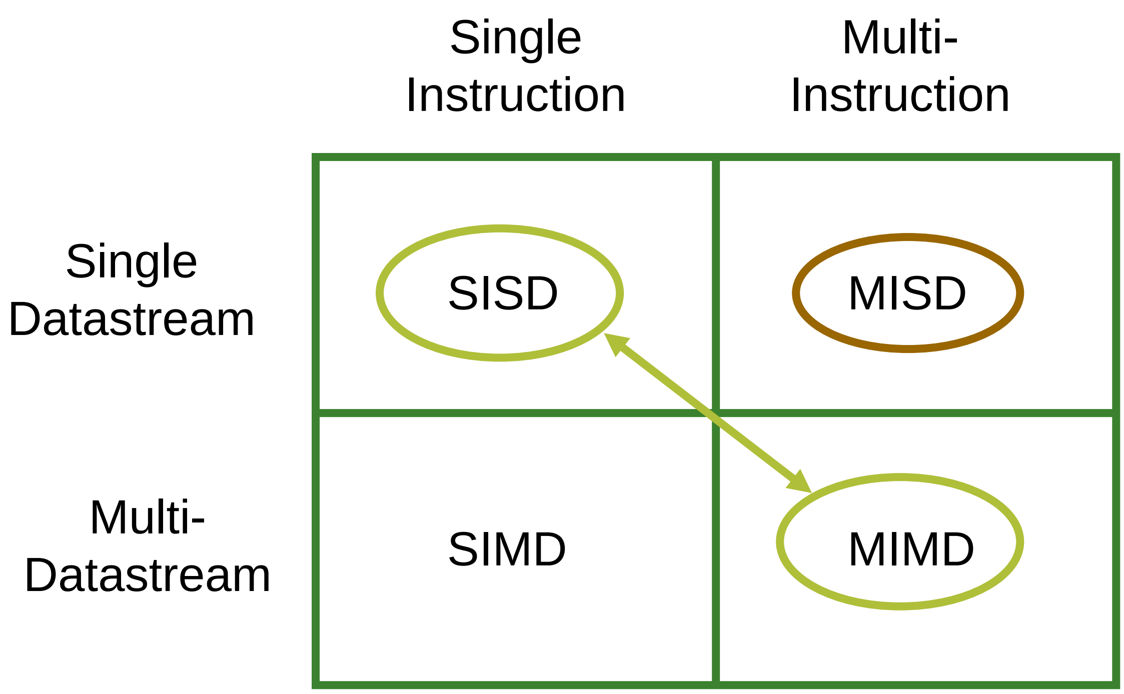
Performance for Heterogeneous Multi-Core Architectures

JoAnn M. Paul

Virginia Tech, ECE
National Capital Region

Common basis for two themes

- Flynn’s Taxonomy
  - Computers viewed from the “inside out”
- Develop a new Taxonomy
  - From the “outside in”
- Later
  - fill in MISD



From the “outside in”

- Single User (SU)
  - a computer designed for use by a single individual at a time
- Multiple User (MU)
  - a computer designed for use by multiple individuals at a time
- Single Application (SA)
  - a computer designed to execute a single application at a time
- Multiple Application (MA)
  - a computer designed to execute multiple applications at a time

The U-A Taxonomy

- Are supercomputers and emerging single chip heterogeneous multiprocessors really the same?

	Single User	Multi-User
Single Application	SUSA	MUSA
Multi-Application	SUMA	MUMA

Filling in the U-A Taxonomy

		SU	MU
SA	SISD	Early Computers, PCs running DOS	Database mainframes (IBM)
	MIMD	Supercomputers	Website servers
MA	SISD	Unix-style O/Ses, Windows	Early timeshare
	MIMD	Emerging! Personal, wireless computers (iPhone++)	General servers (Google)

Implications of SUMA-MIMD

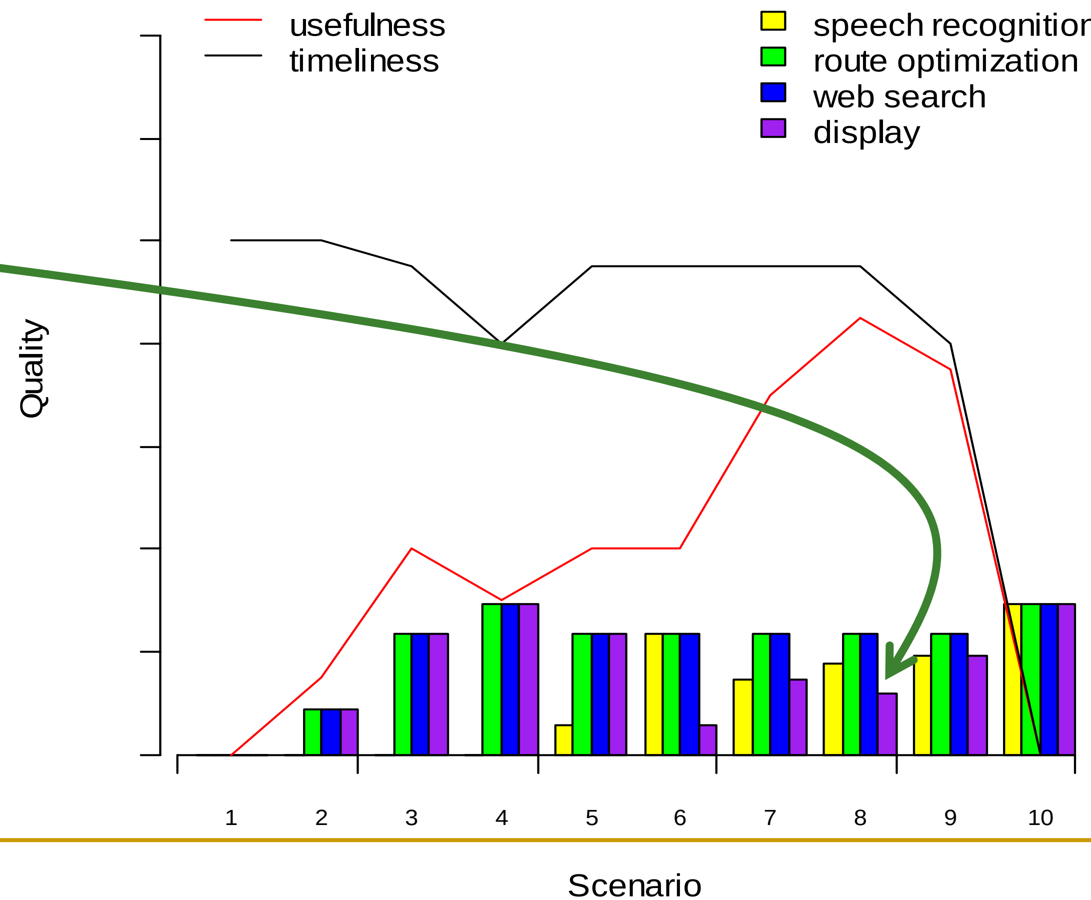
- Performance is judged solely by the user
  - Can even be highly *personalized*
  - Diminishing returns for
    - Graphics, sound quality
- Users can only juggle so many things at once
  - Will never see some performance improvements
  - Will trade more apps for isolated optimality

Usefulness and Timeliness

- Addition of speech to a navigation system

Scenario	Description
1	no software
2	basic function
3	best without speech
4	RISC load too high
5	inadequate speech
6	inadequate display
7	speech and display equal
8	optimal balance
9	SIMD load too high
10	all load too high

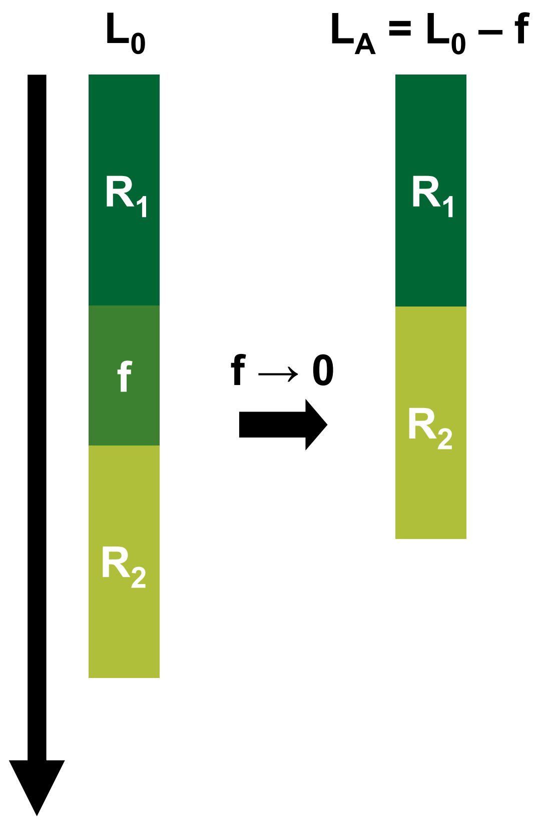
Is slower better?



Amdahl's Law

- Speedup
  - Limited to the removal of the sequential fraction
    - Ideal if it can be considered to
      take zero time to execute,
    - you can't do better than

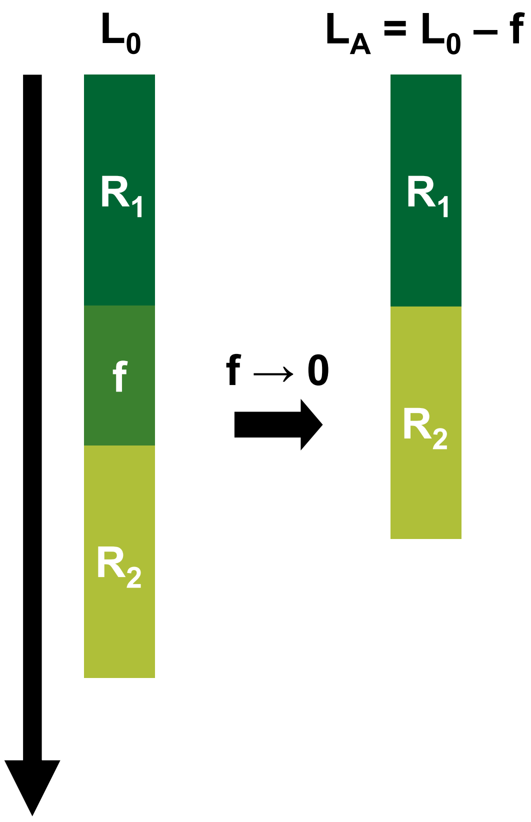
$$S = \frac{R_1 + f + R_2}{R_1 + R_2}$$



Amdahl's Law

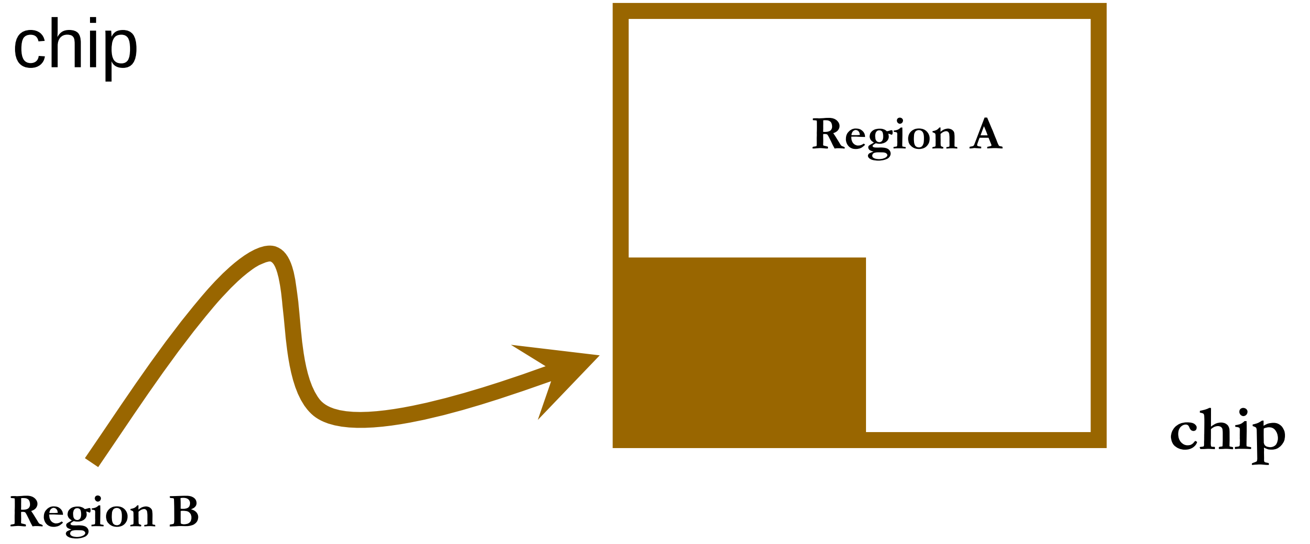
- Implies
  - slower *f* is worse
- Assumes
  - Other regions are unaffected
  - Independence of design concerns

$$S = \frac{R_1 + f + R_2}{R_1 + R_2}$$



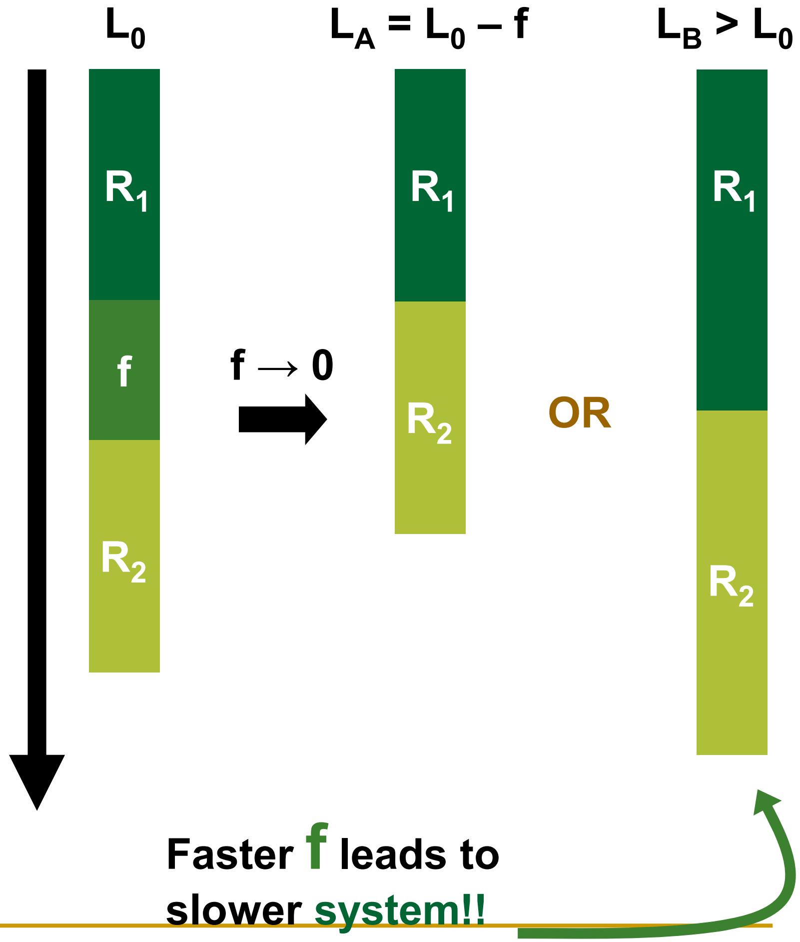
Global Effects

- Anything with a common element
  - what does not have a common element
- Easiest to see is a boundary
  - e.g. chip

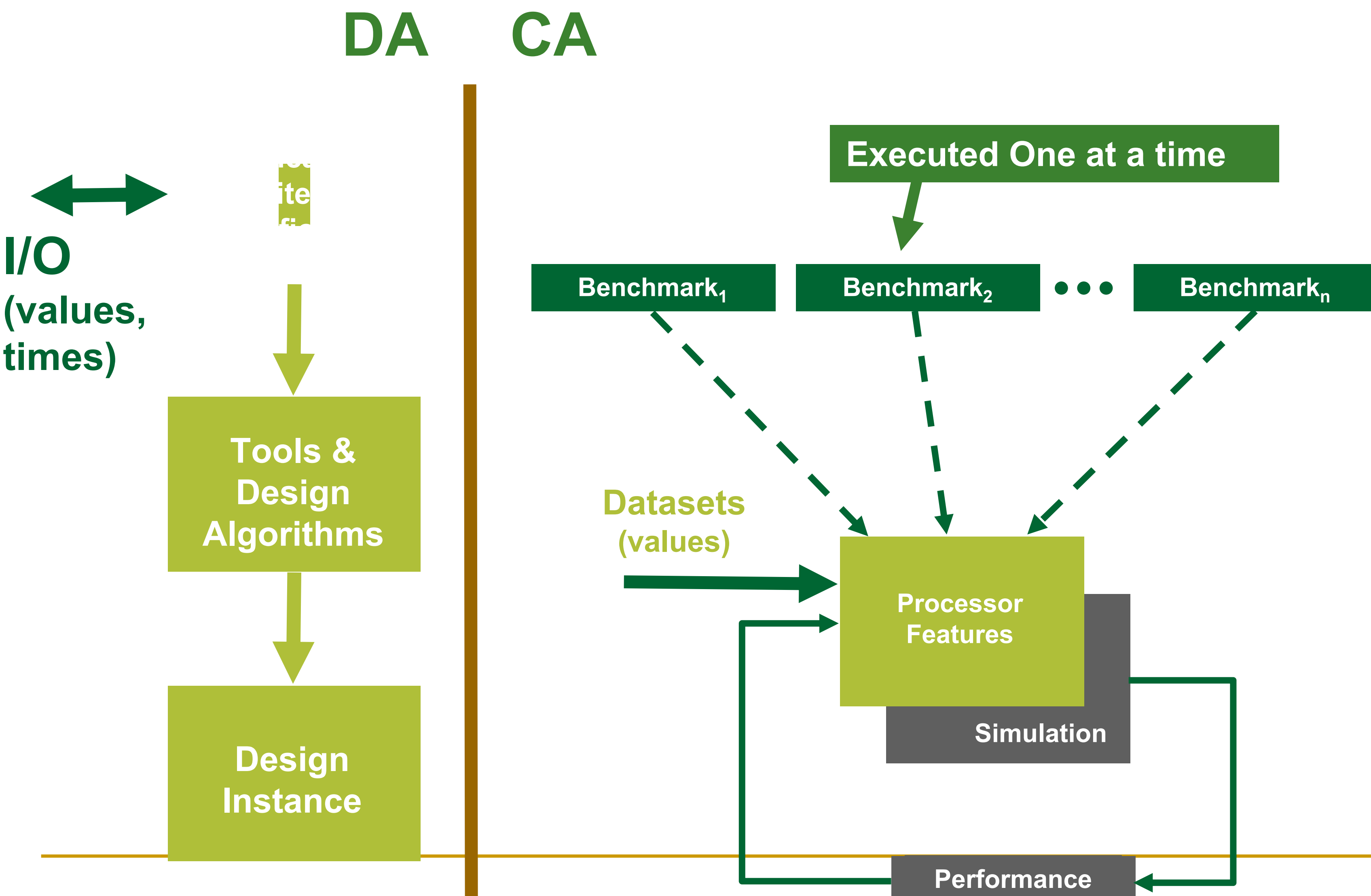


When Bounded, Heterogeneous

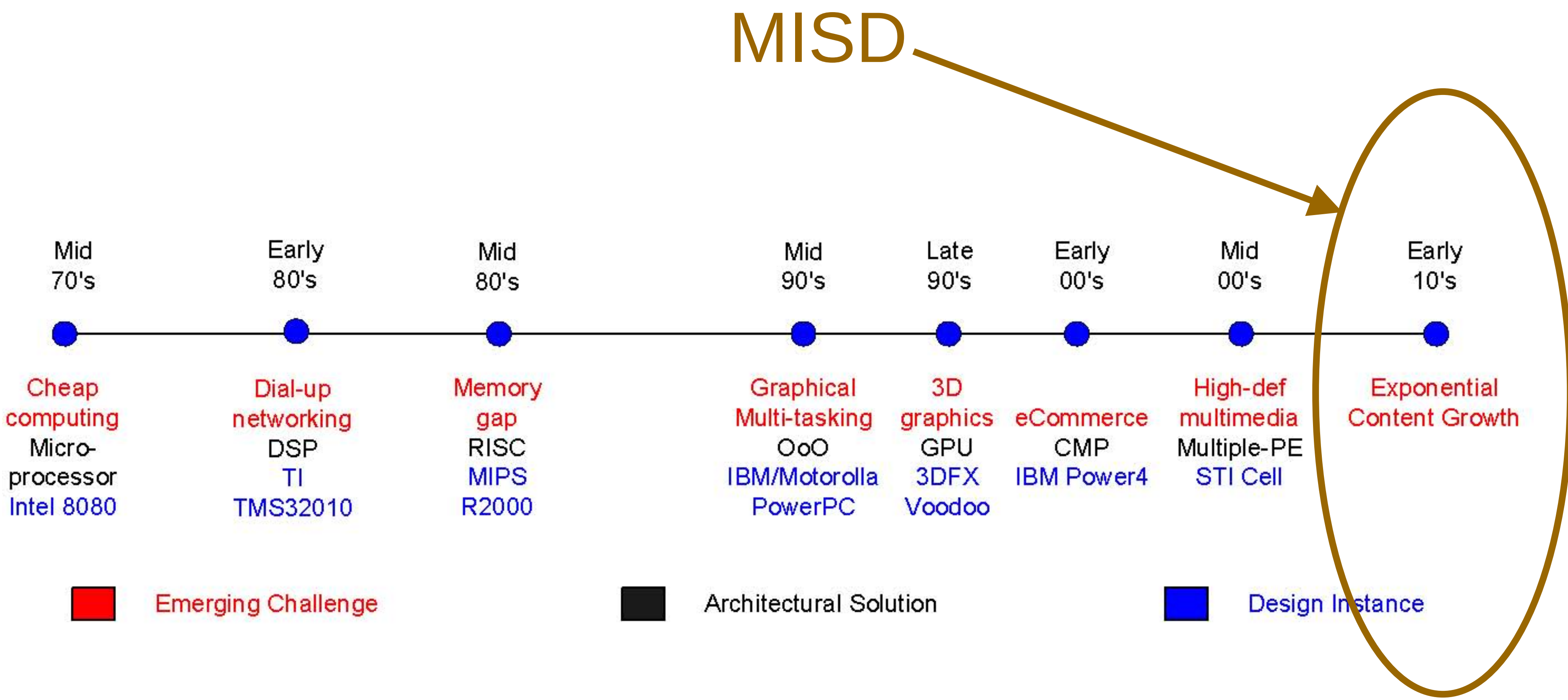
- The system can be faster when the isolated portion (sequential fraction) is slowed down
  - Microarchitecture has this effect also
  - Floating point vs. larger register file



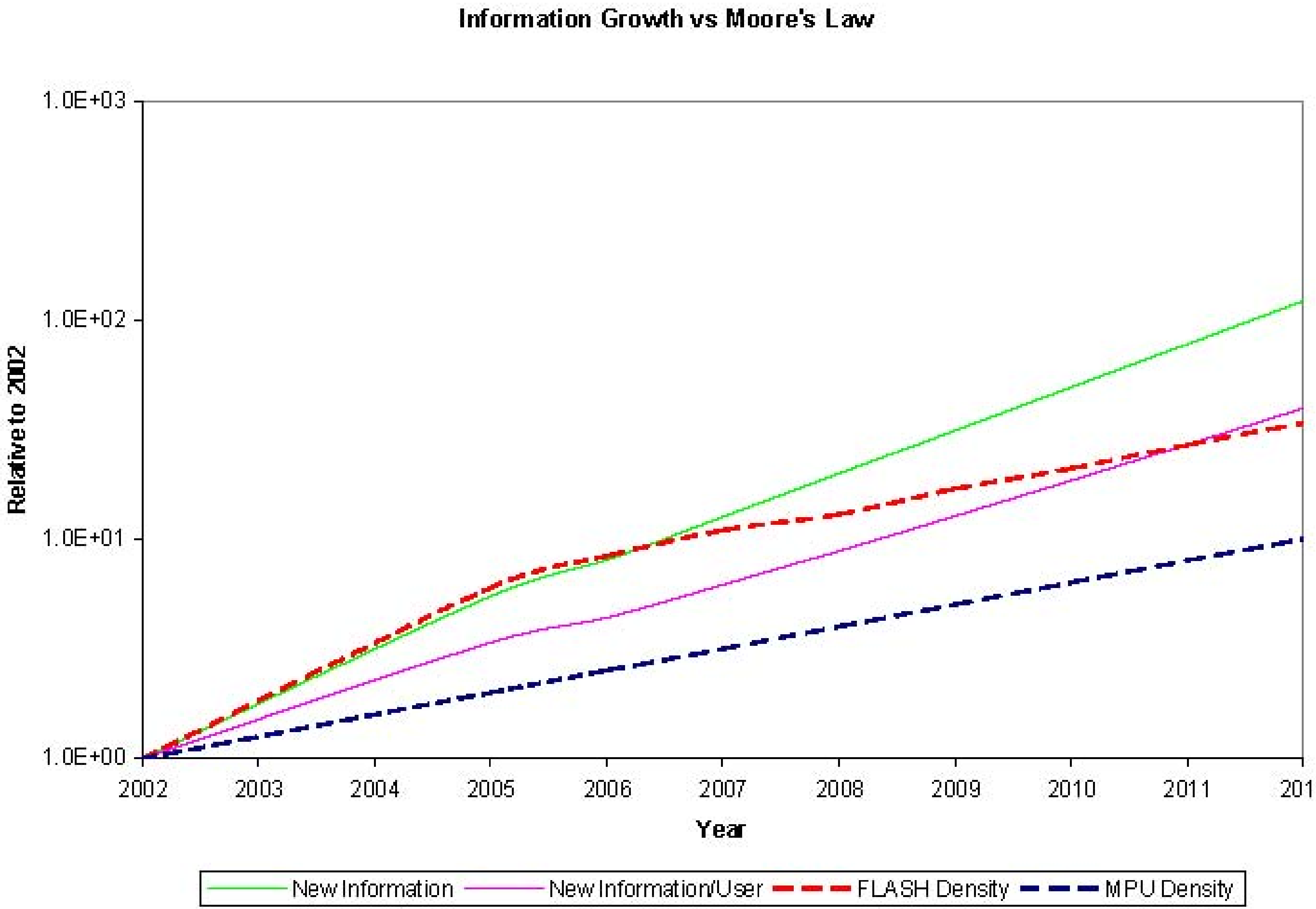
Faster is no longer always better



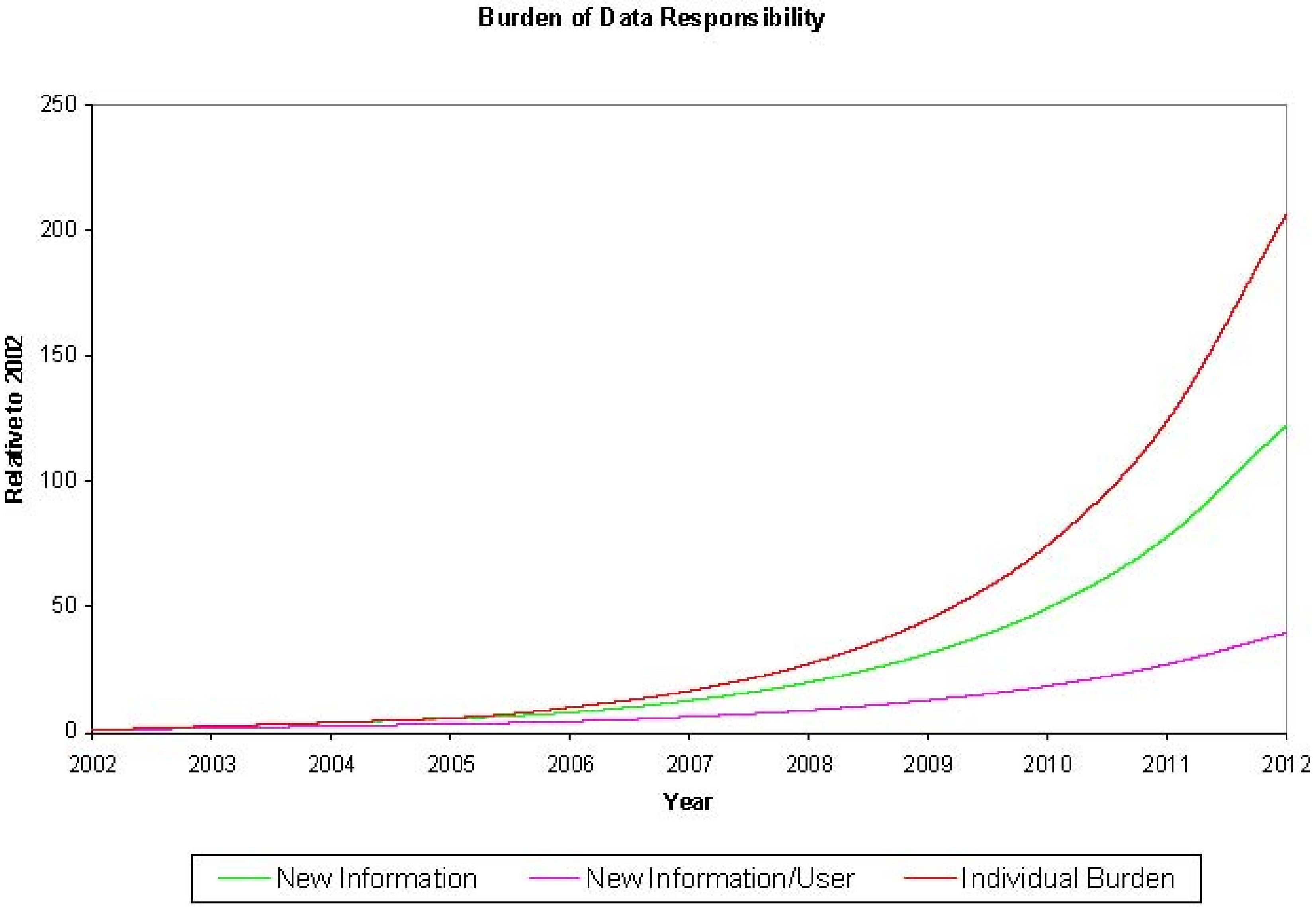
Challenges and Architectural Responses



Information v. transistor density growth



Effort required to find, file and share data



Computing has been Quantitative

- Transformative
  - mapping from one mathematical space, A, to another, B, functionally, completely for all a ε A and b ε B
    - b = F(a)
- Functional transformations
  - Tend to be reductive
    - typically transforming points from spaces with more to those with fewer dimensions
  - Are objective
    - Mathematical – the “correct” result is presumed to exist

Qualitative Computing

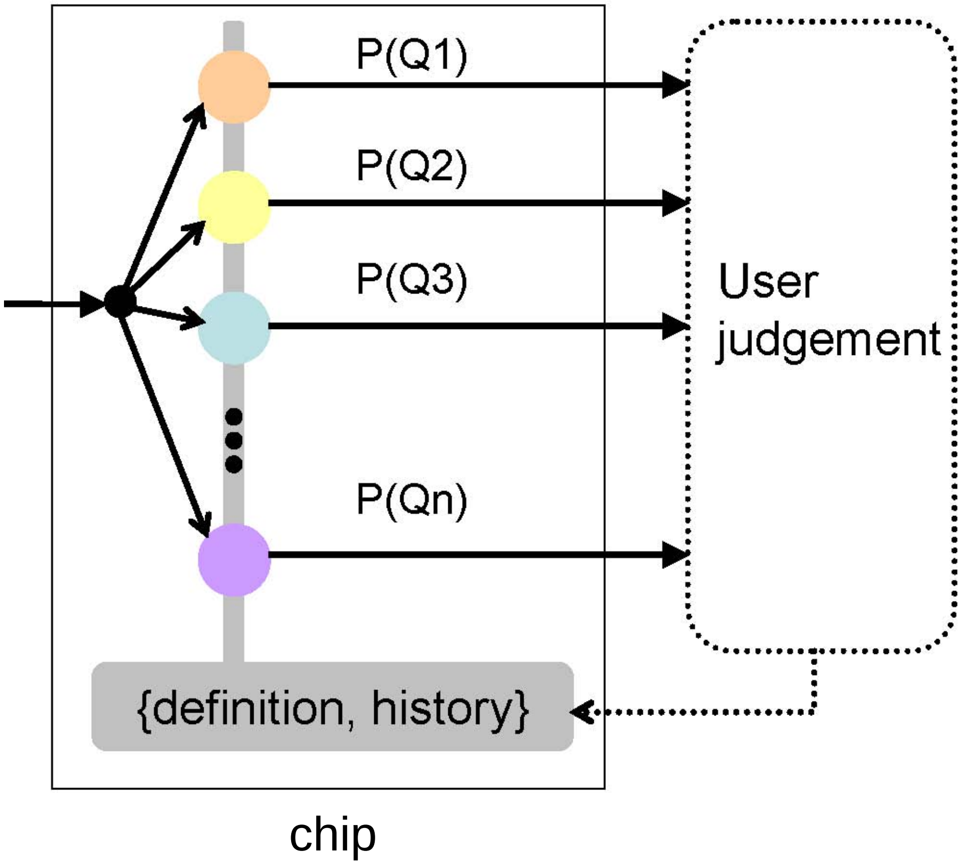
- Identify
  - The extent to which some input has a set of qualities {A’, B’, C’, D’, …}
    - where those qualities are represented in imperfect mathematical spaces
- Subjective
  - Real answer only lies in the interpretation of *individual users*
  - Exact interpretations do not exist
    - When is something, “red, large, good” ?

Solution is Architectural, MISD

- Personalization
  - tuning to individual preferences
    - enabled by data management to on personalized, (and mobile) computers
- Mult-interpretation
  - simultaneously applying different techniques
    - enabled by heterogeneity of the underlying architecture
- Integration
  - effective coupling of global data and processing
    - enabled by new data-processor relationships on novel *single chip heterogeneous multiprocessor architectures*

Spectroprocessing is MISD

- For Qualitative Computation
  - Customized to
    - personal preference
    - history
  - Grows over time
    - May start with common seed
  - Facilitates judgment



Conclusions

- Since the dawn of computing
  - Performance has been driven by
    - Reducing latency
    - Increasing throughput
  - All computation has been quantitative
    - fundamentally transformational
- We are at the frontier of transcending each