System-Level Analysis and Design under Uncertainty

Ivan Ukhov

December 15, 2017

Introduction

Electronic Systems

- Omniscient
- Omnipresent

Analysis and Design

- Challenging
- Consequential

Uncertainty

- Lack of knowledge
- Inherent randomness

Fabrication

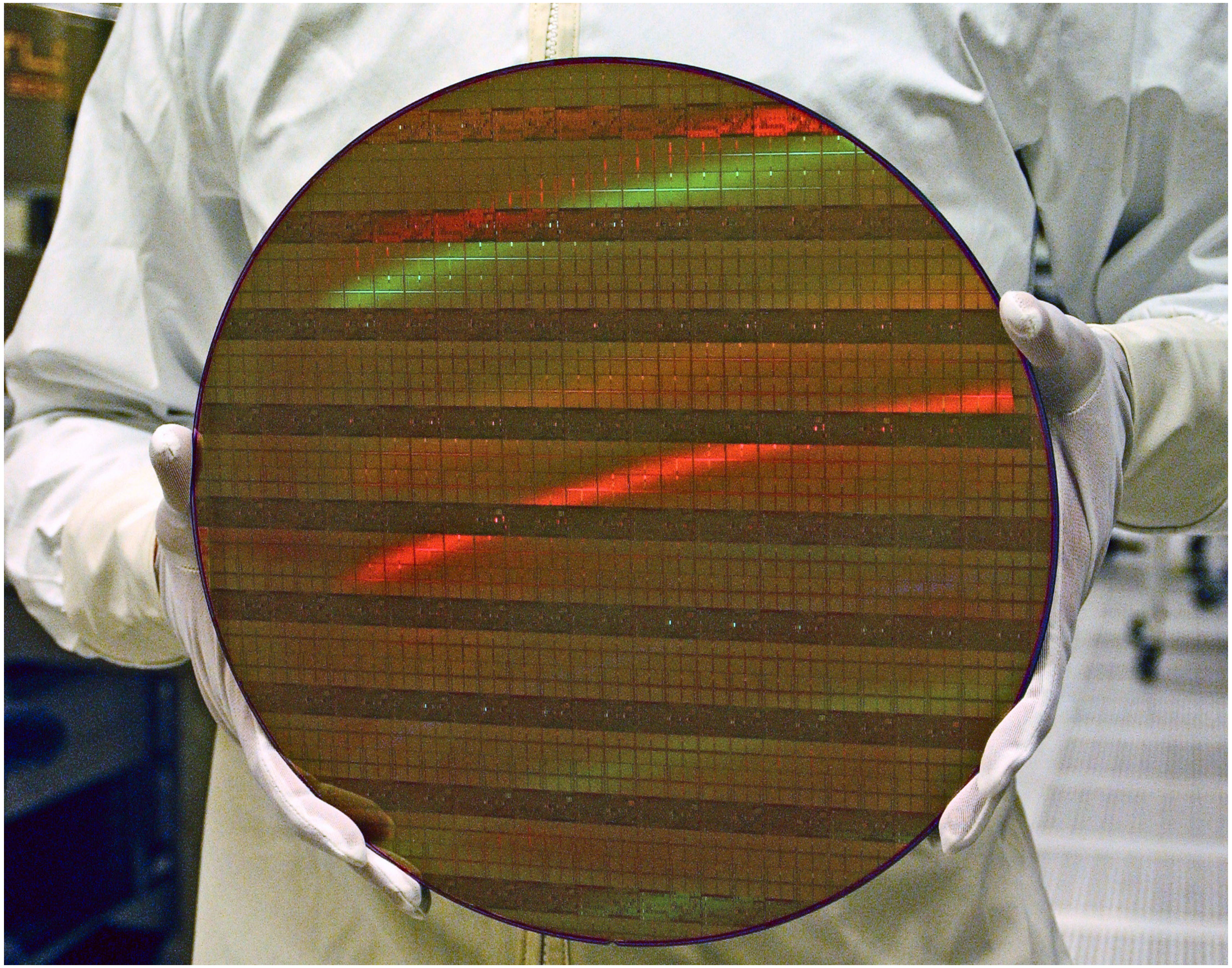
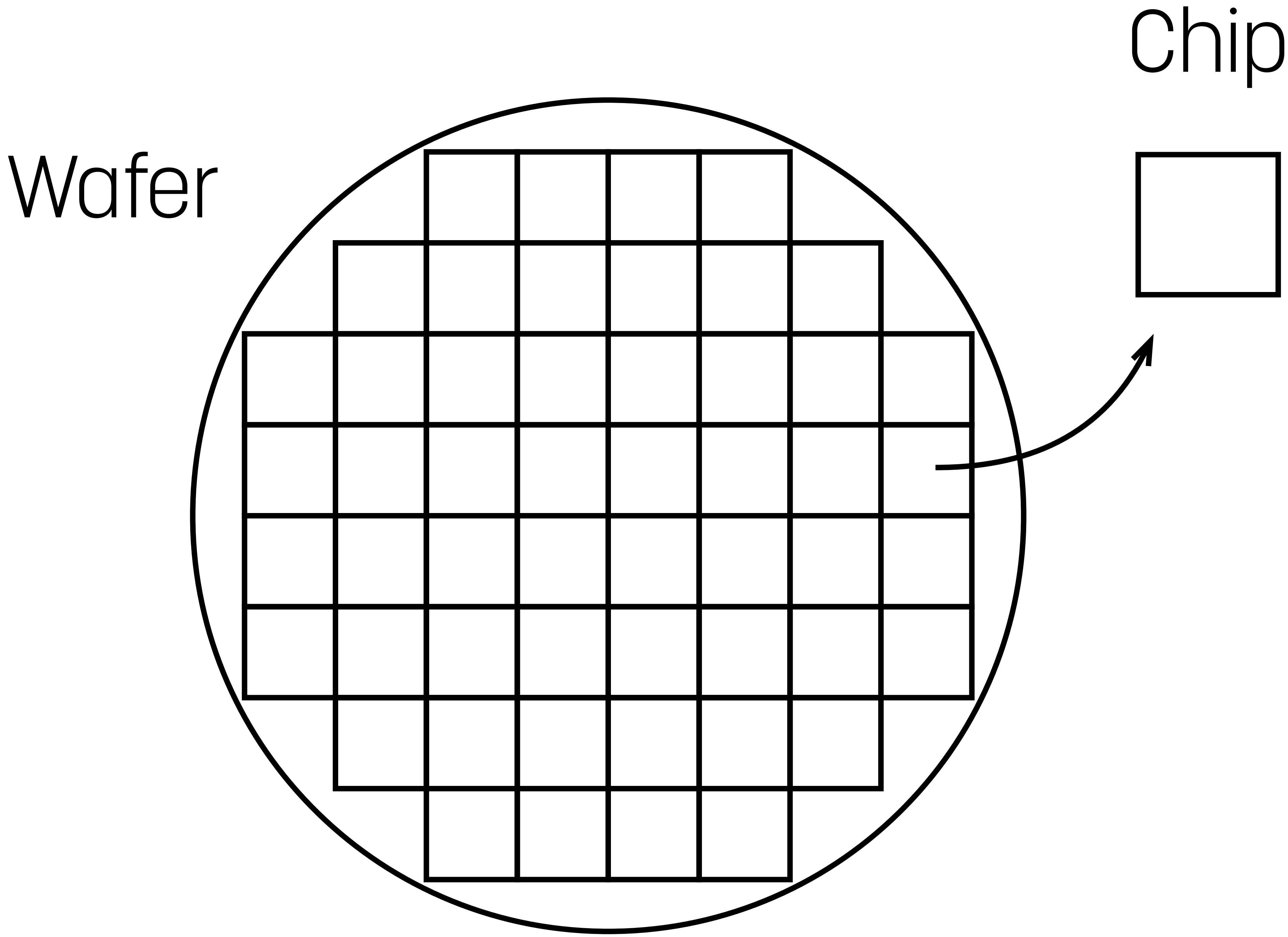
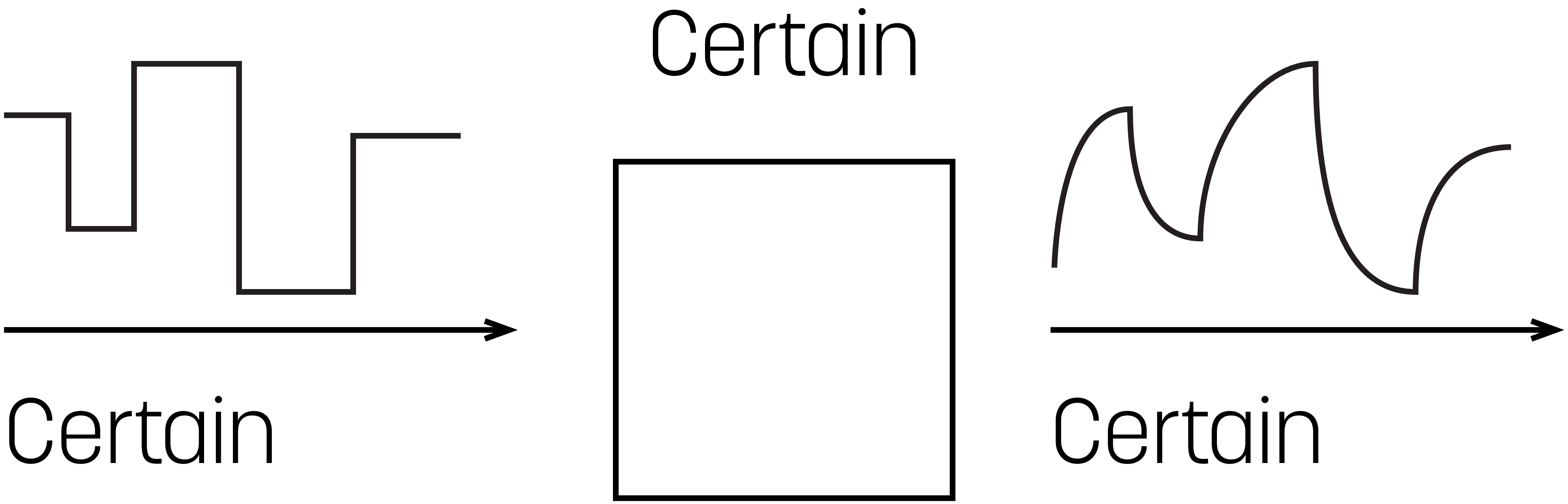


Photo: Intel, https://goo.gl/mHMxe1

Fabrication

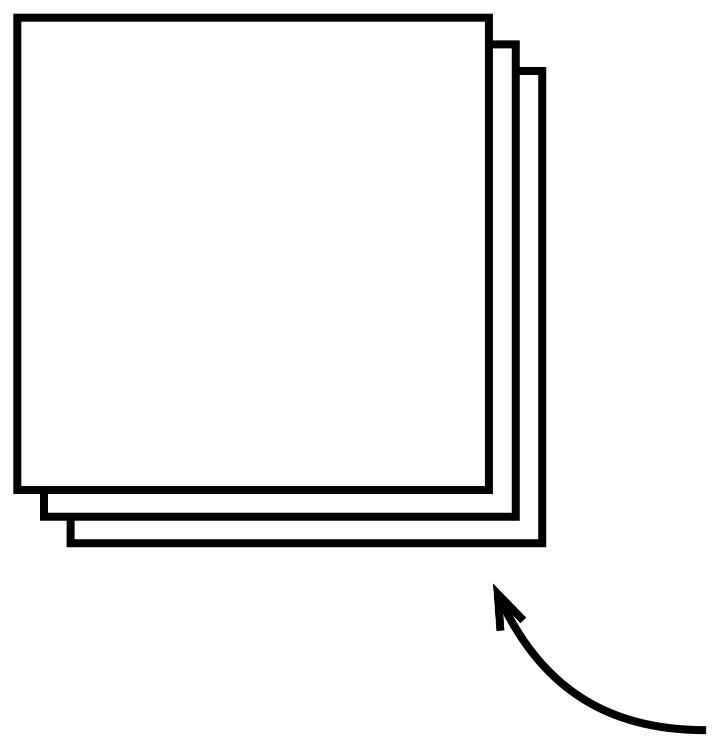


No Variation



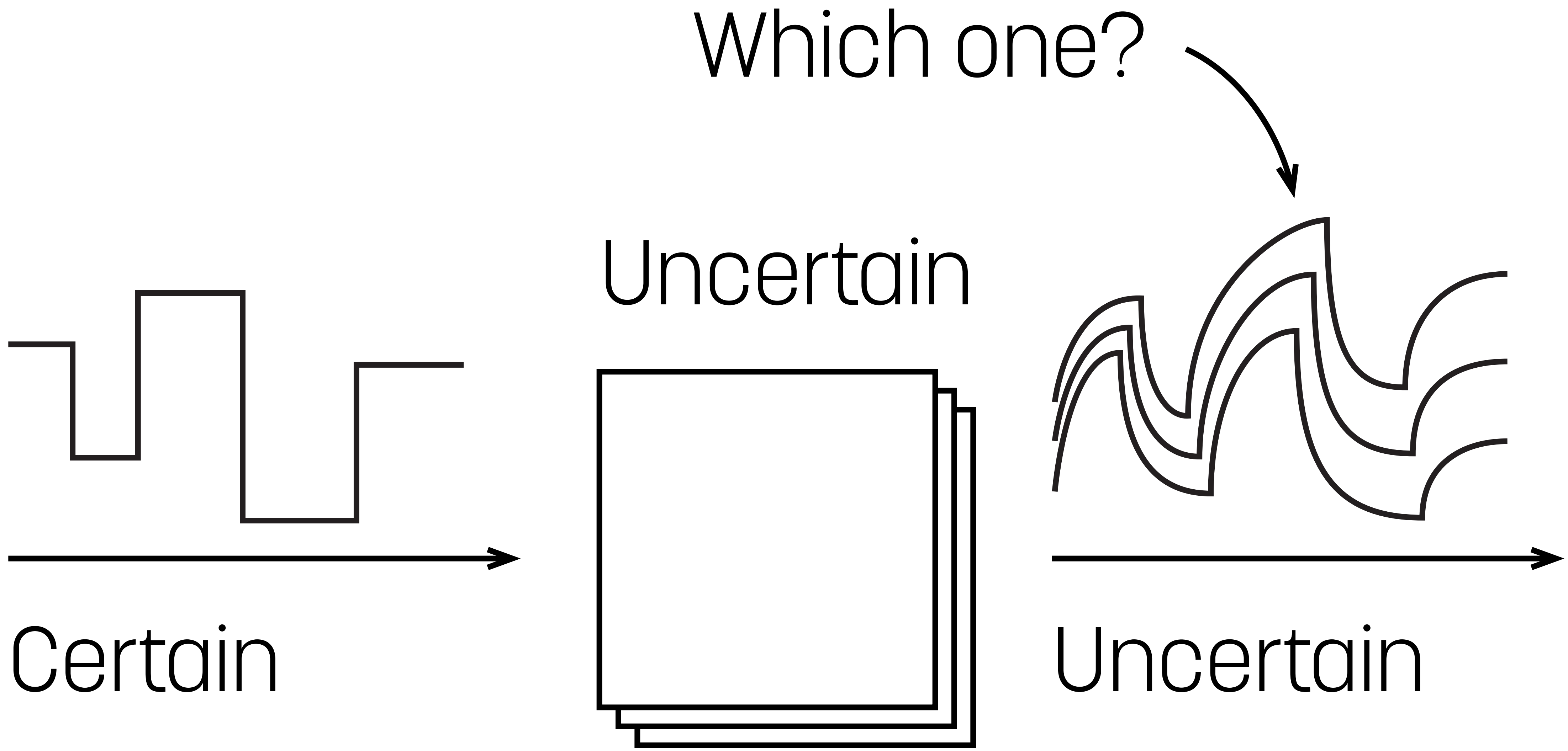
Process Variation

Uncertain

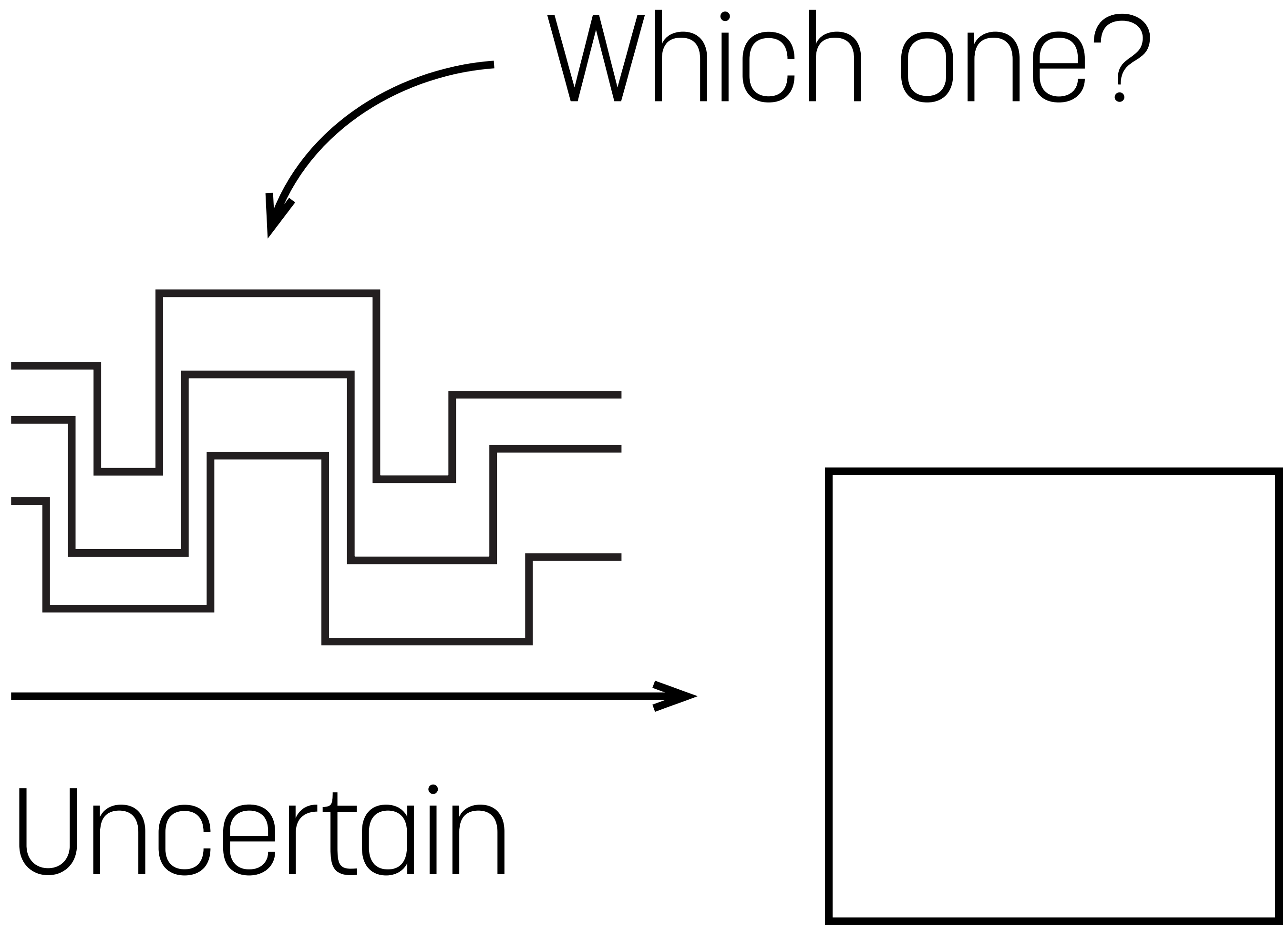


Which one?

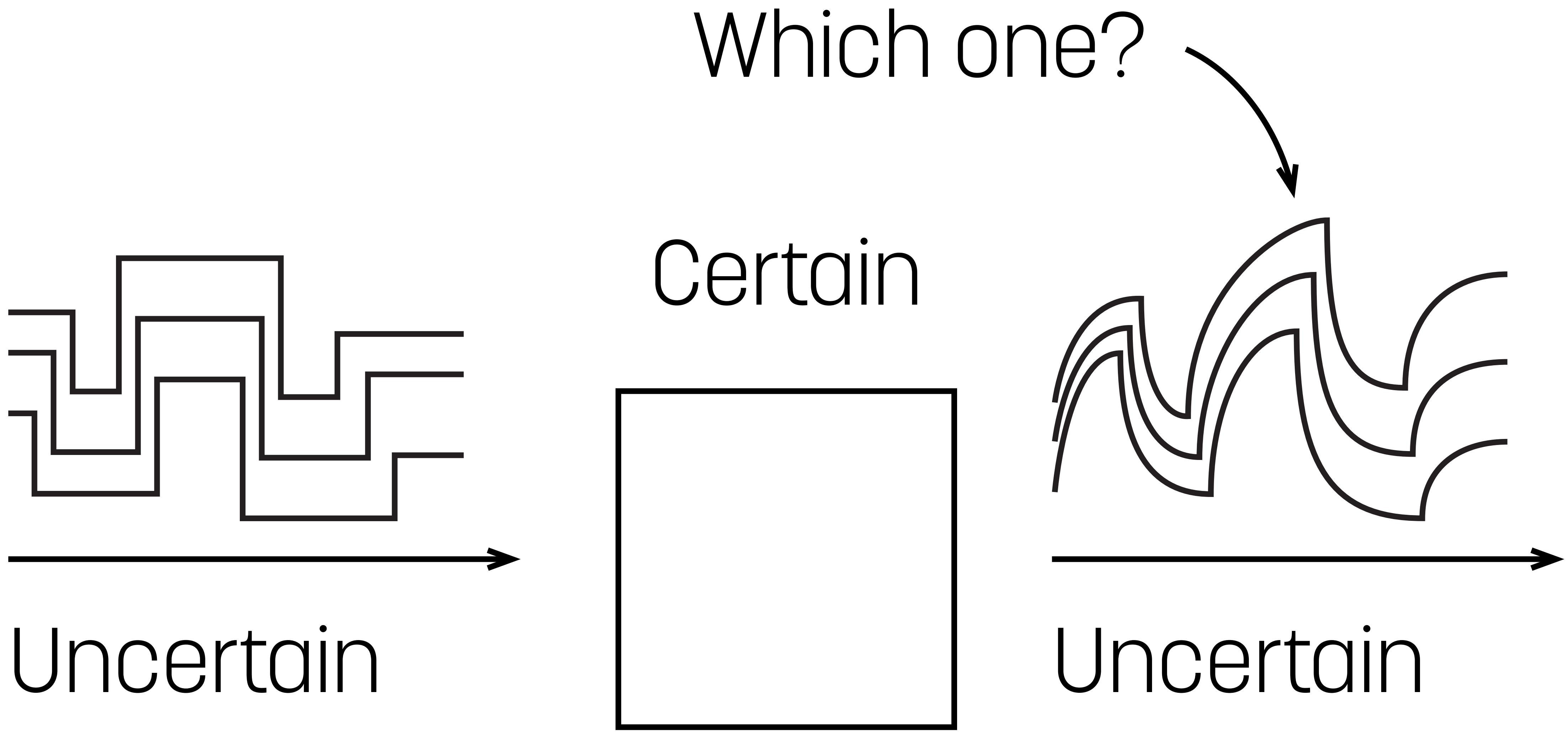
Process Variation



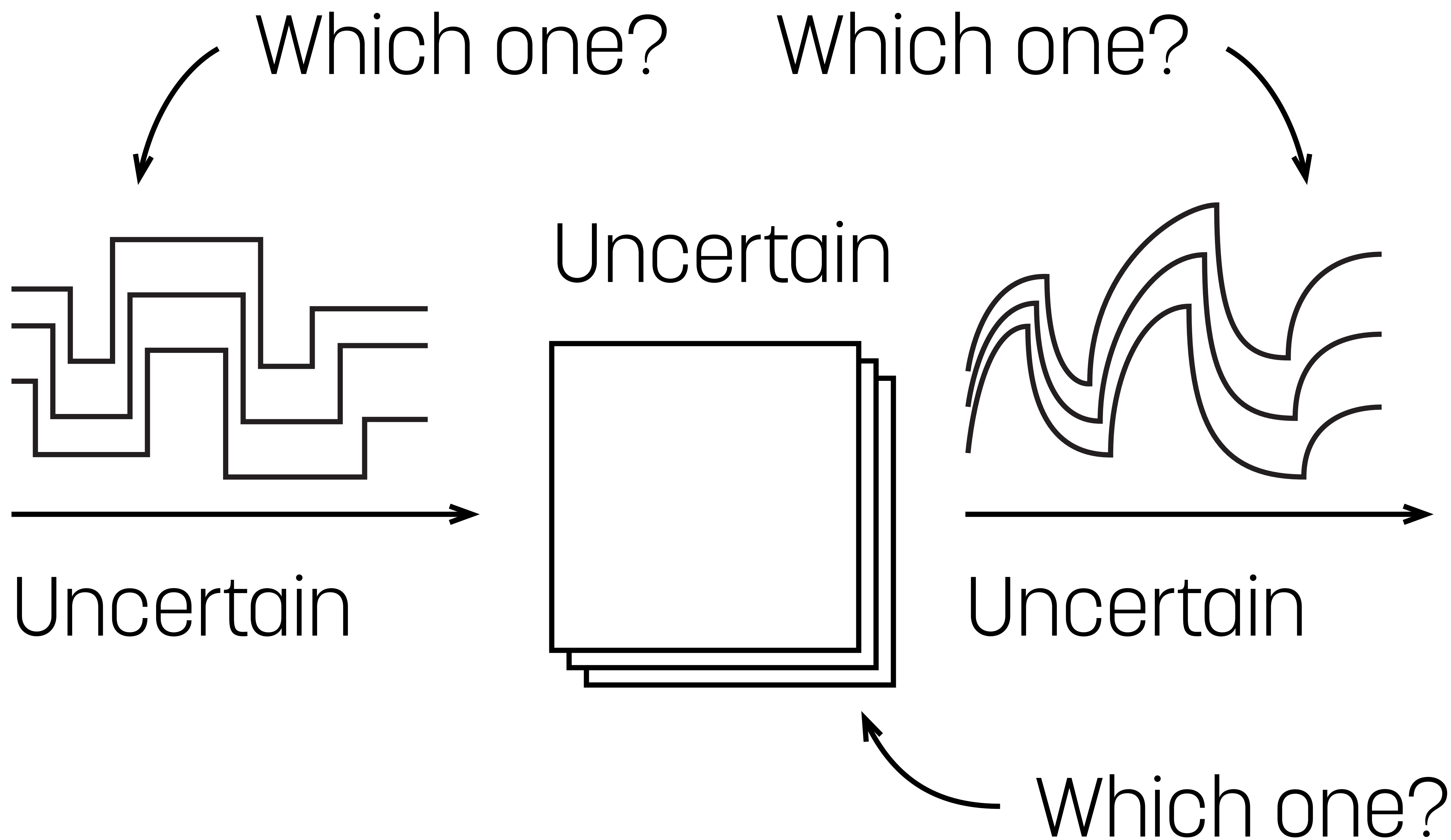
Workload Variation



Workload Variation



Analysis and Design



Motivation

- Uncertainty
  - Inevitable
  - Deleterious

Objective

- Provide the designer with effective and efficient techniques for analysis and design under uncertainty

Outline

1. Analysis and Design with Certainty
2. Characterization of Process Variation
3. Analysis and Design under Process Variation
4. Analysis under Workload Variation
5. Resource Management under Workload Variation

Analysis and Design with Certainty

Power and Temperature

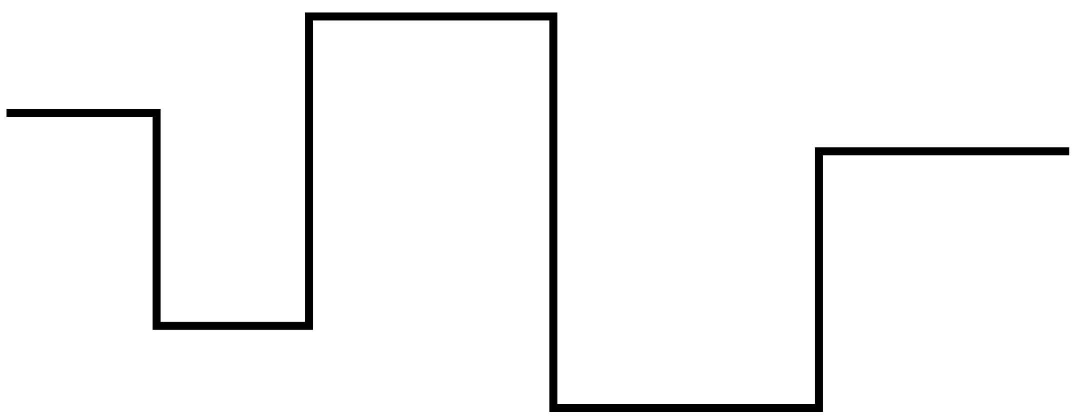
- Highly important
  - Energy efficiency
  - Reliability

Temperature Analysis

- Transient state
- Dynamic steady state

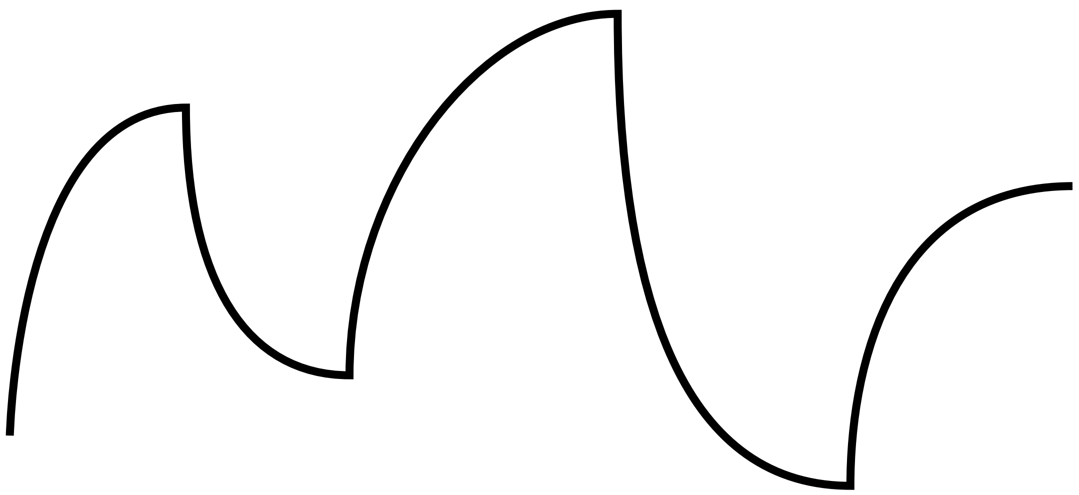
Transient State

Power



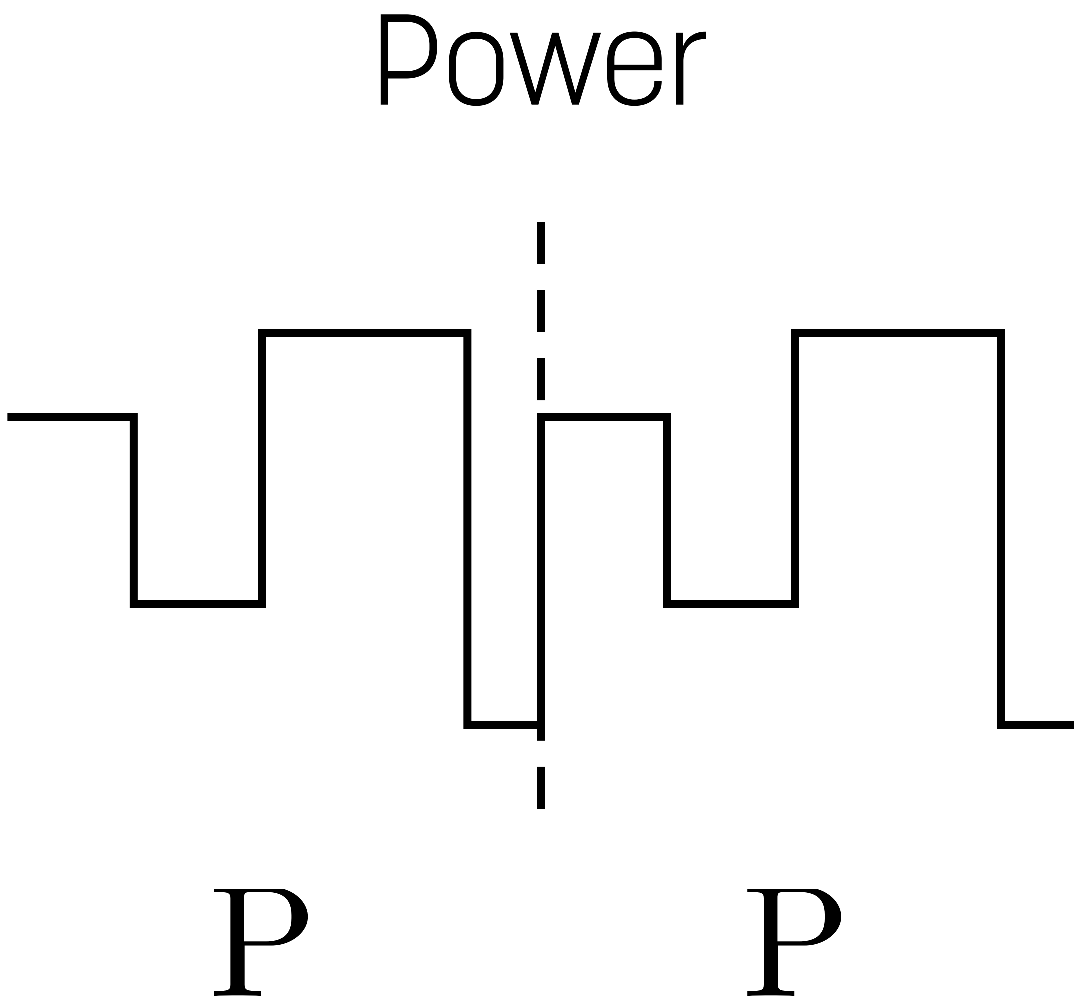
P

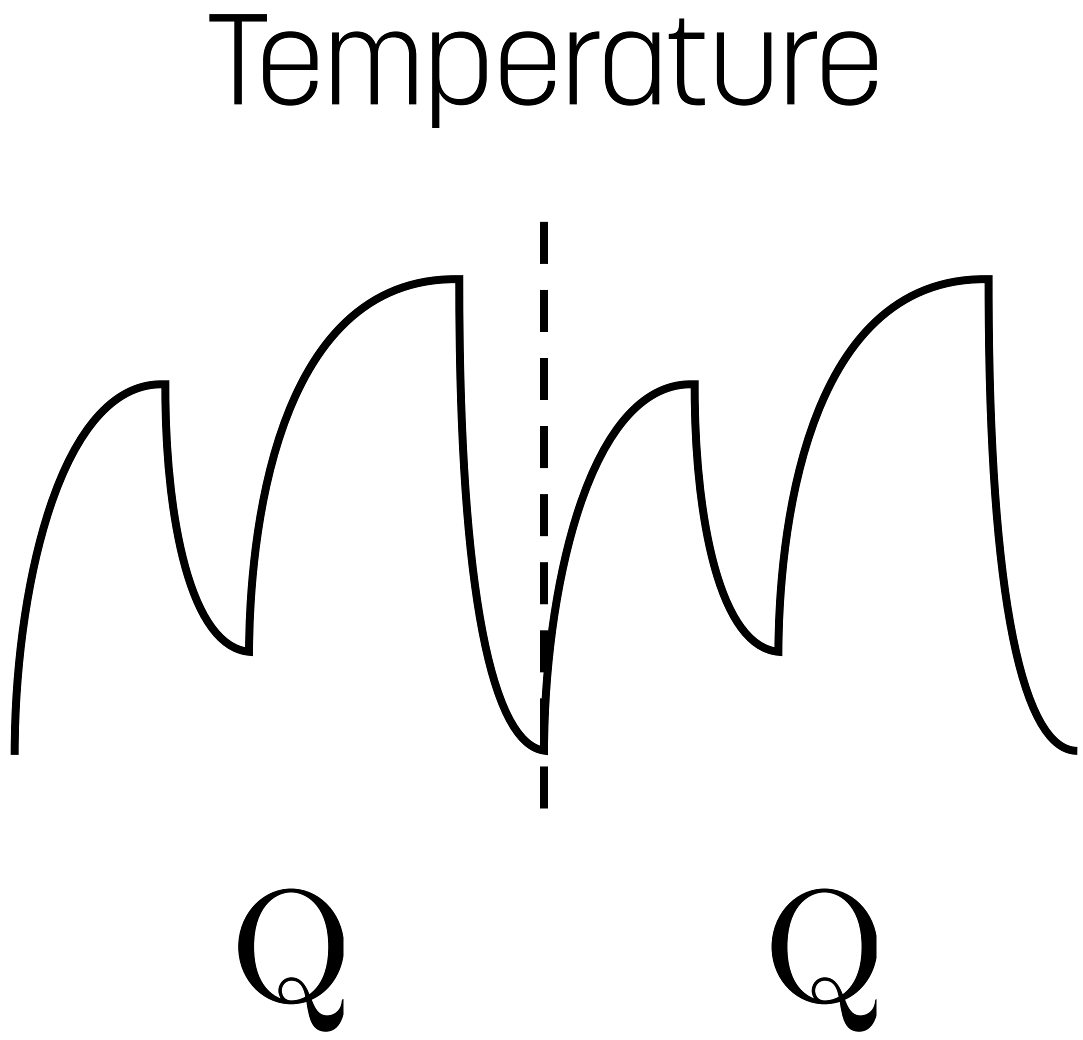
Temperature



Q

Dynamic Steady State





Problem Formulation

- Given a periodic power profile P
- Compute the corresponding dynamic steady-state temperature profile Q

Previous Work

- Slow
- Inaccurate

Proposed Solution

- Fast
- Exact

Solution Overview

Thermal RC model

Power

$$\begin{cases} \mathrm{C}\dfrac{\mathrm{d}\tilde{s}(t)}{\mathrm{d}t} + \mathrm{G}\tilde{s}(t) = \tilde{\mathrm{B}}p(t) \\ q(t) = \tilde{\mathrm{B}}^{\mathrm{T}}\tilde{s}(t) + q_{\mathrm{amb}} \end{cases}$$

Temperature

Solution Overview

Auxiliary transformation

$$s(t) = \mathrm{C}^{\frac{1}{2}}\,\tilde{s}(t)$$

$$\mathrm{A} = -\mathrm{C}^{-\frac{1}{2}}\mathrm{G}\mathrm{C}^{-\frac{1}{2}}$$

$$\mathrm{B} = \mathrm{C}^{-\frac{1}{2}}\tilde{\mathrm{B}}$$

$$\begin{cases} \dfrac{\mathrm{d}s(t)}{\mathrm{d}t} = \mathrm{A}s(t) + \mathrm{B}p(t) \\ q(t) = \mathrm{B}^{\mathrm{T}}s(t) + q_{\mathrm{amb}} \end{cases}$$

Solution Overview

Dynamic steady state

$$s_0 = s_{n_s}$$

$$w_0 = 0$$

$$w_i = \mathrm{E}w_{i-1} + \mathrm{F}p_i$$

$$s_0 = \mathrm{U}(\mathrm{I} - e^{\Lambda\tau})^{-1}\mathrm{U}^{\mathrm{T}}w_{n_s}$$

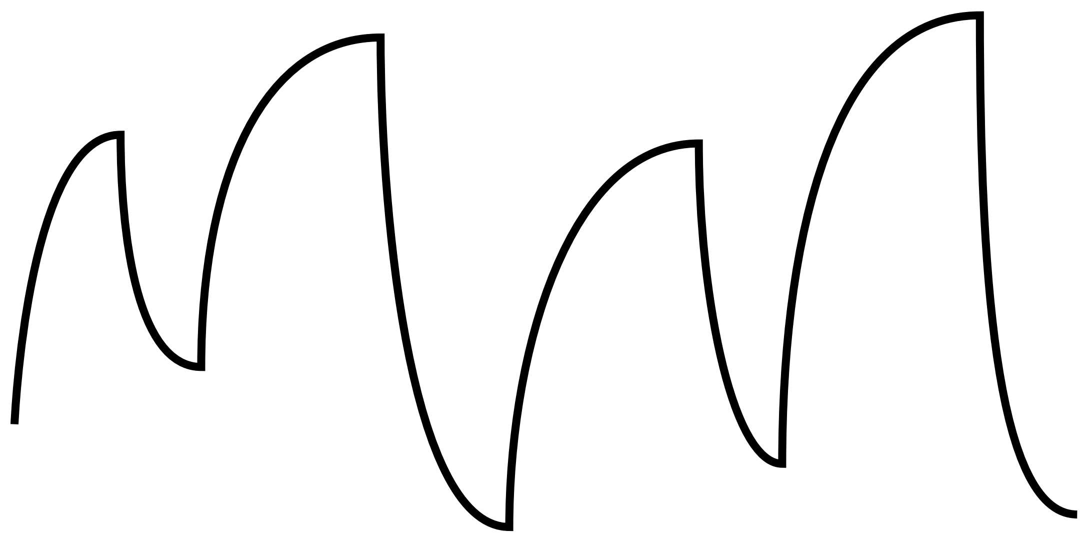
$$s_i = \mathrm{E}s_{i-1} + \mathrm{F}p_i$$

Experimental Results

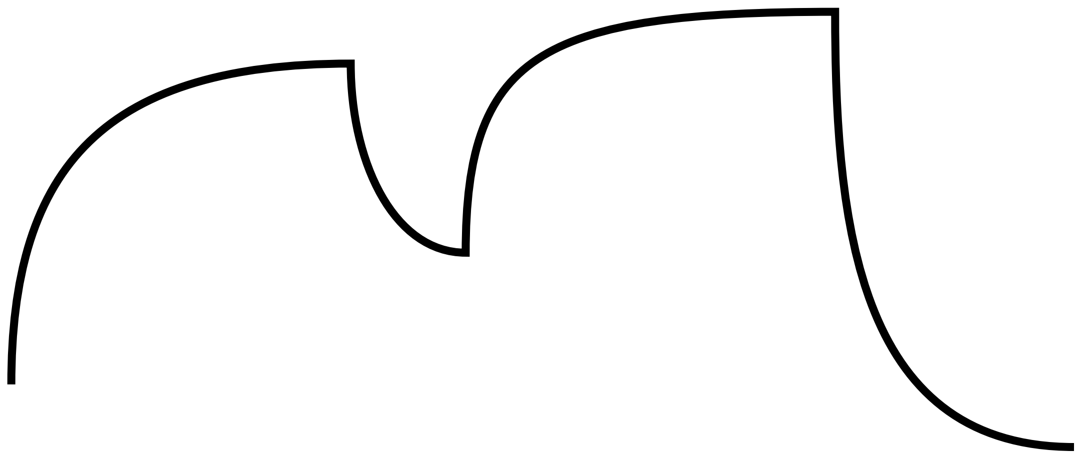
- Considered diverse scenarios
- Shown high computational speed
  - 9-170 times faster than analytical iterative transient analysis
  - 2000-5000 times faster than iterative analysis with HotSpot

Thermal Cycling

More damage



Less damage



Reliability Optimization

- Vary the application's schedule
- Maximize the system's lifetime
- Satisfy a number of constraints

Reliability Optimization

Mean time to failure

$$\max_{\mathcal{S}} \min_{i=1}^{n_p} \mu_i(\mathcal{S})$$

Schedule

$$\tau(\mathcal{S}) \leq \tau_{\max}$$

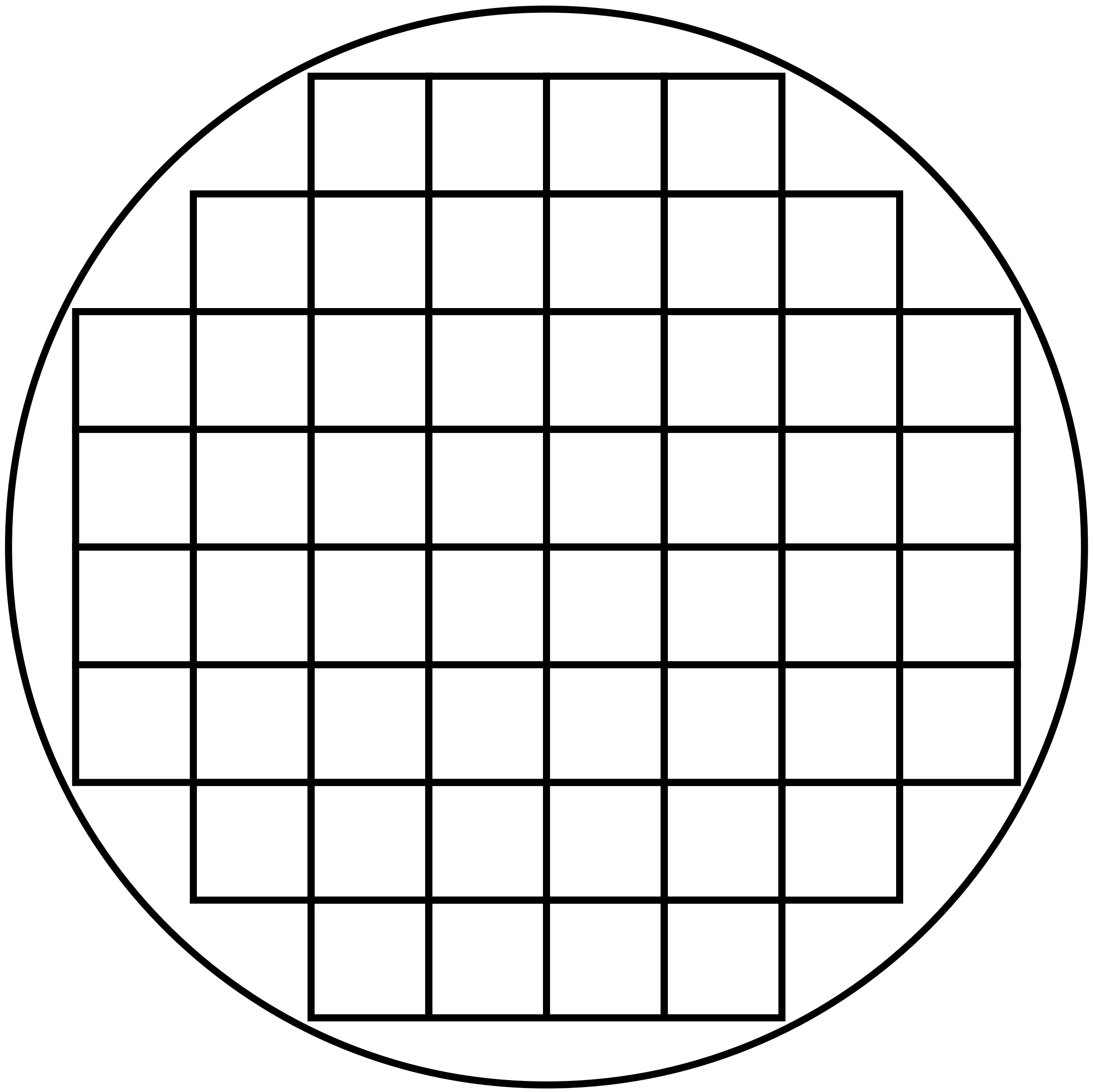
$$Q(\mathcal{S}) \leq q_{\max}$$

Experimental Results

- Applied to a large set of synthetic problems and to a real-life problem
- Increased the mean time to failure by a factor of 10–70
- Maintained energy efficiency

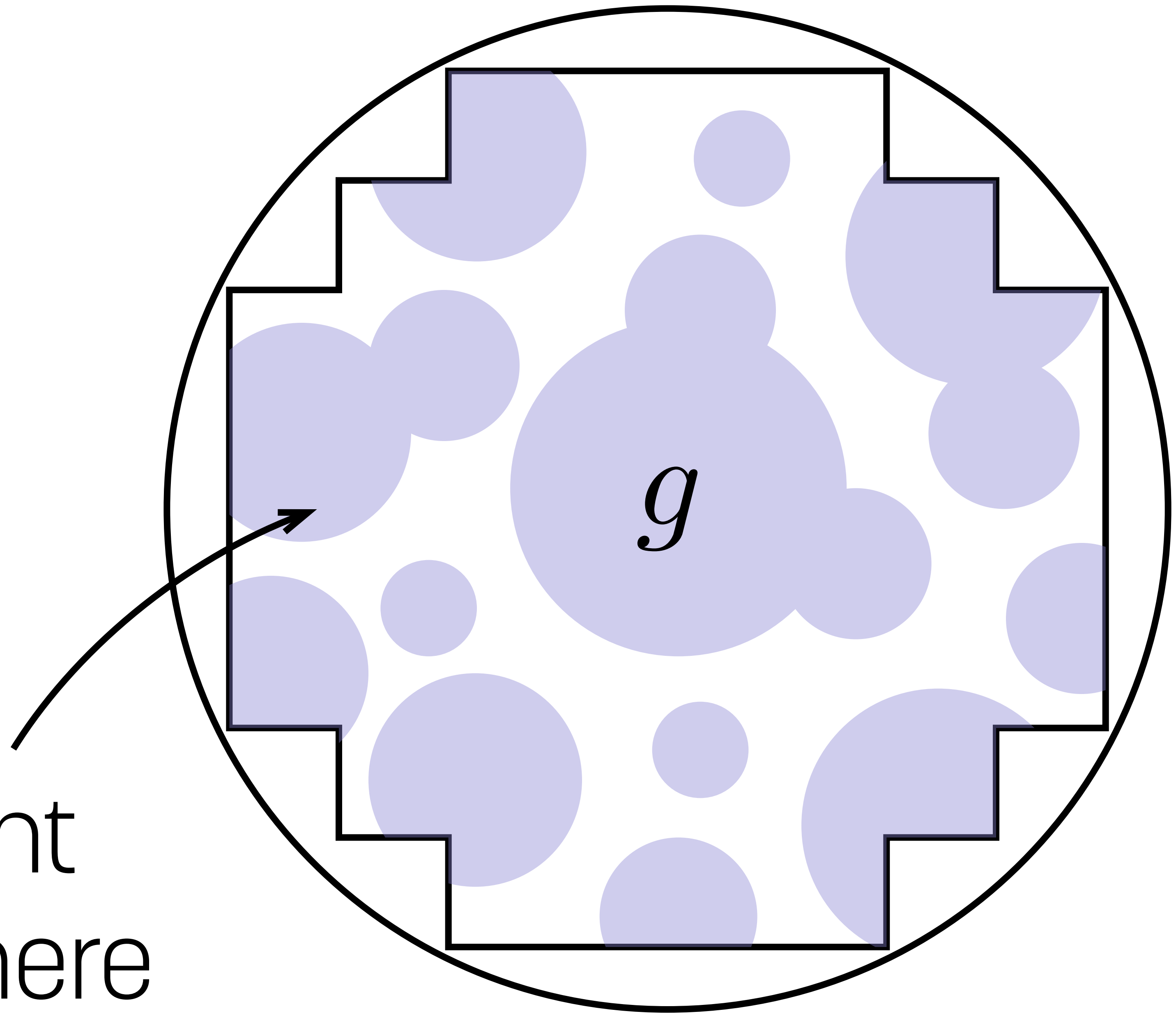
Characterization of Process Variation

Fabrication



Quantity of Interest

Uncertain



Different everywhere

Problem Formulation

- Given the knowledge of the technological process at hand
- Quantify the process parameter g at all locations on the wafer

Previous Work

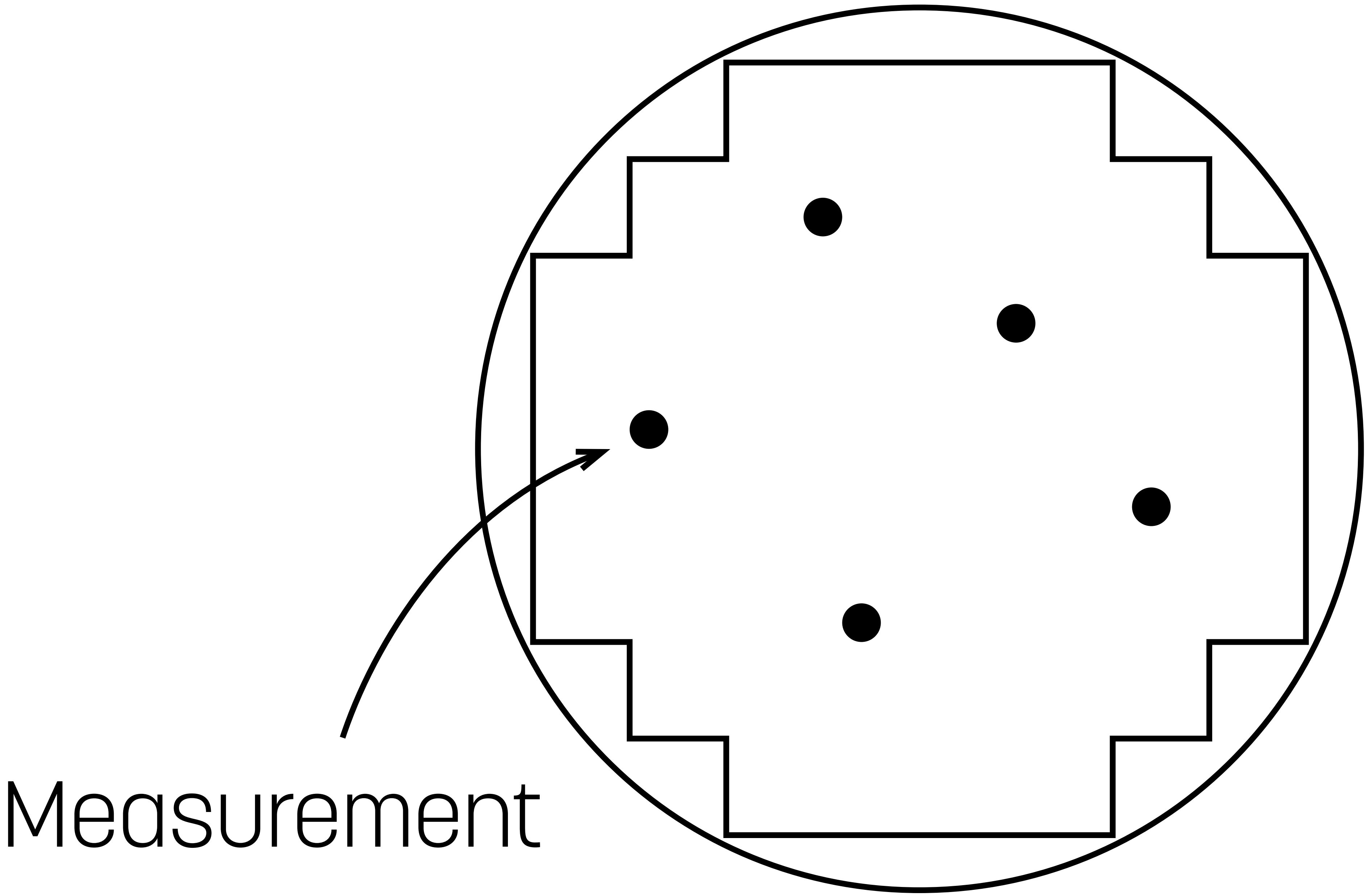
- Intrusive
- Secondary parameters

Proposed Solution

- Non-intrusive
- Primary parameters

Solution Overview

$$h = f(g)$$



Solution Overview

$$g|u \sim \text{Gaussian Process}(\mu, v)$$

$$\epsilon|u \sim \text{Gaussian}\left(0, \sigma_\epsilon^2\right)$$

Noise

Bayes' theorem

$$p(u|H) \propto p(H|u)p(u)$$

Solution Overview

Metropolis-Hastings algorithm

$$u \sim t_{n_u}\left(u^*, \alpha^2 \mathrm{J}^{-1}\right)$$

Posterior optimization

$$u^* = \arg\max_u p(u|H)$$

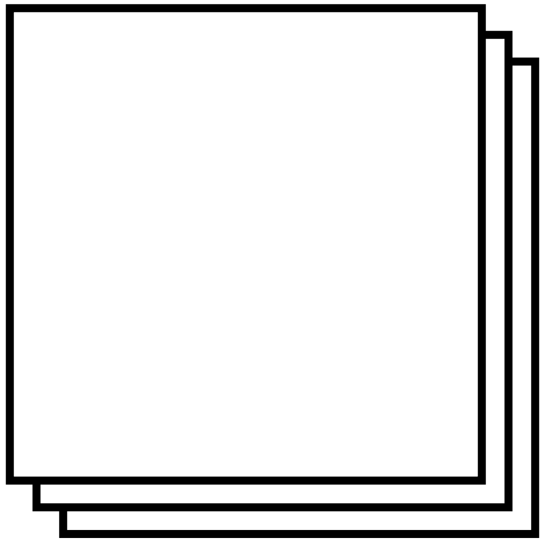
Experimental Results

- Inferred the effective channel length from temperature measurements
- Considered diverse configurations
- Shown high accuracy and speed
  - Less than 5% of error
  - Less than 20 minutes

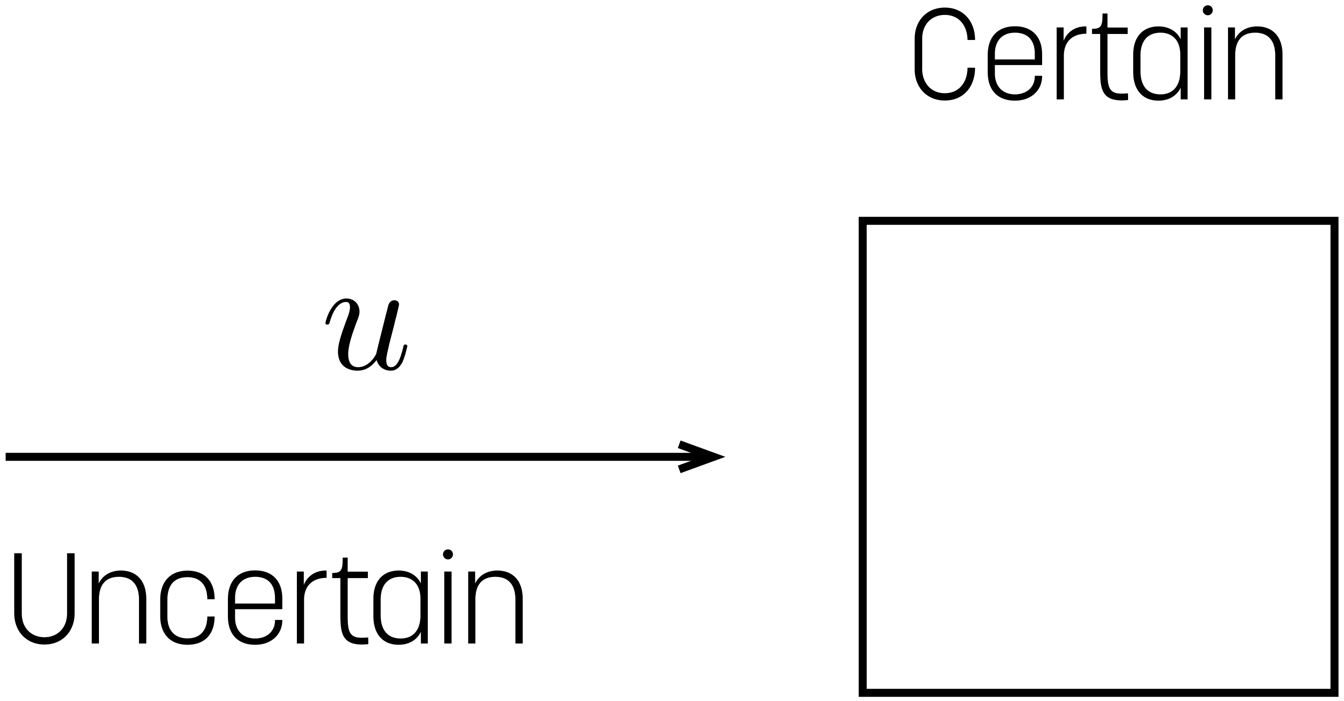
Analysis and Design under Process Variation

Process Variation

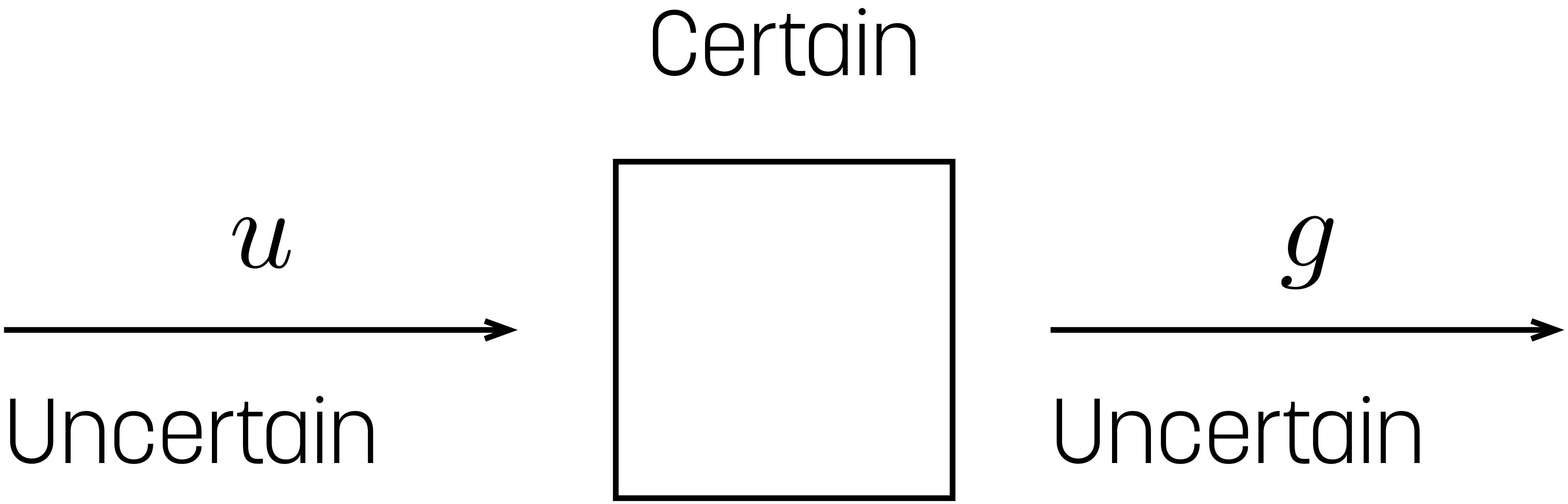
Uncertain



Uncertain Parameters



Quantity of Interest



Problem Formulation

- Given the probability distribution of the uncertain parameters u
- Compute the probability distribution of the quantity of interest g

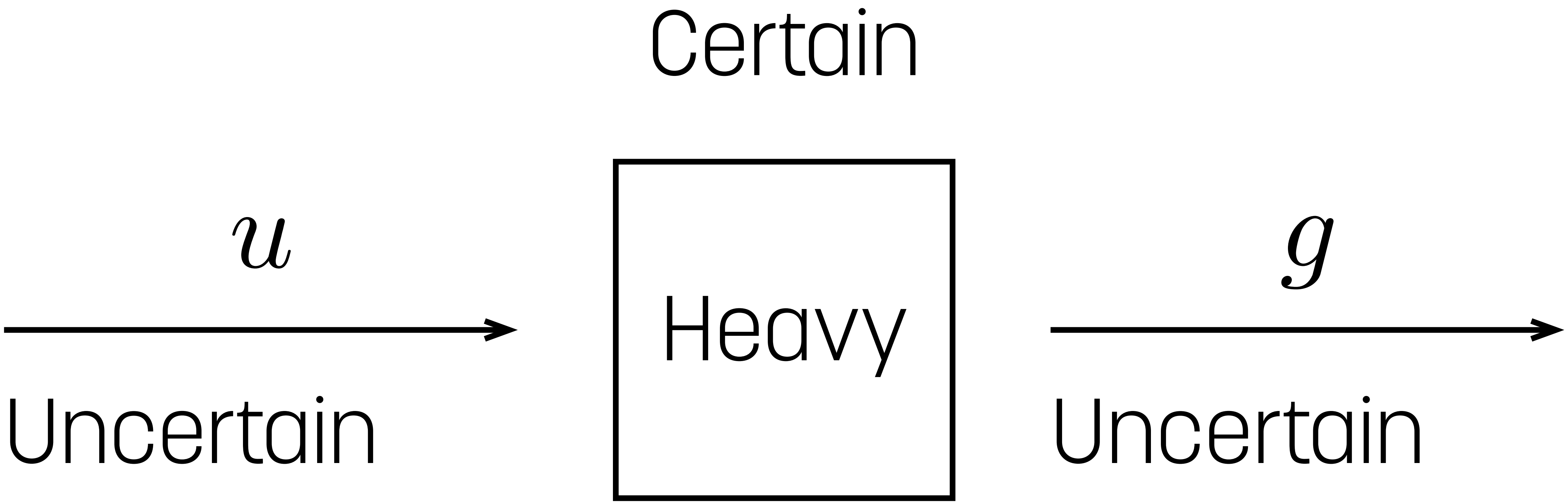
Previous Work

- Limited to specific quantities
- Unrealistic assumptions

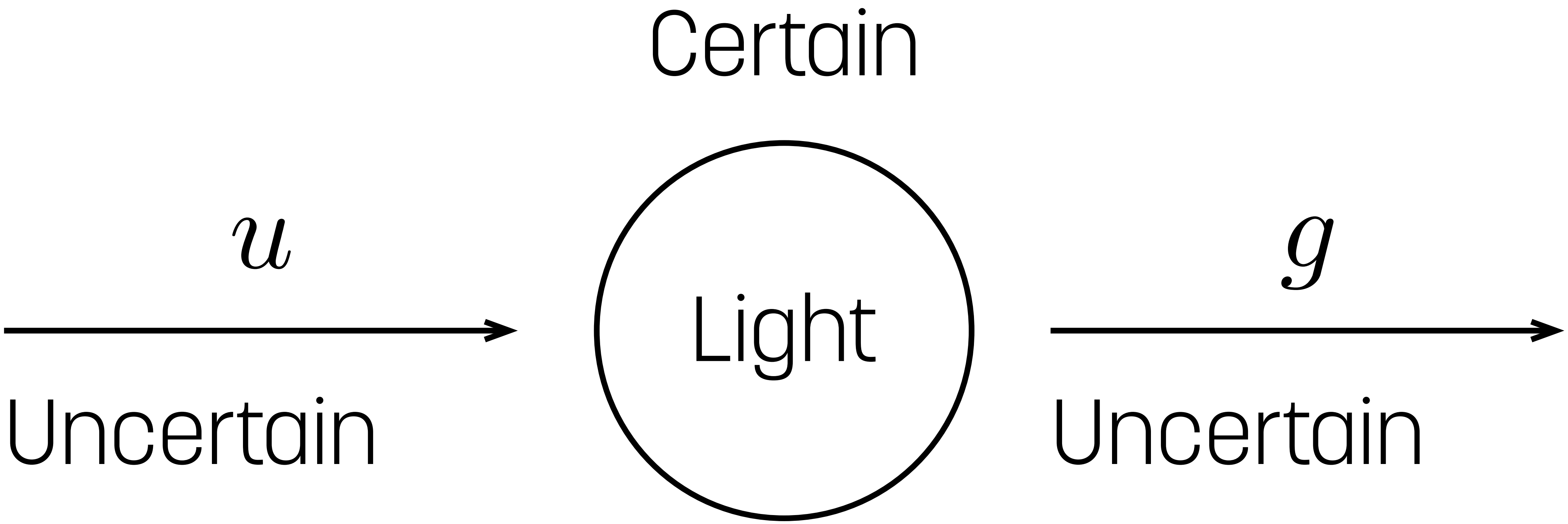
Proposed Solution

- General
- Efficient
- Easy to apply

Solution Overview



Solution Overview



Solution Overview

$$u = \mathbb{T}(z)$$

$u : \Omega \to \mathbb{R}^{n_u}$ Many dependent

$z : \Omega \to \mathbb{R}^{n_z}$ A few independent

$$g(u) = (g \circ \mathbb{T})(z) = g(\mathbb{T}(z))$$

Solution Overview

Polynomial chaos

$$g \approx \mathcal{C}_{l_c}^{n_z}(g) = \sum_{i \in \mathcal{I}_{l_c}^{n_z}} \hat{g}_i \psi_i$$

Polynomial

Spectral projection

$$\hat{g}_i \approx \mathcal{Q}_{l_q}^{n_z}(g\psi_i) = \sum_{j \in \mathcal{J}_{l_q}^{n_z}} (g \circ \mathbb{T})(z_j)\psi_i(z_j)w_j$$

Coefficient

Temperature Analysis

- Account for process variation
  - Transient state
  - Dynamic steady state

Experimental Results

- Compared with extensive simulations
- Shown high accuracy and speed
  - Less than 2% of error
  - 3–5 orders of magnitude faster than direct sampling

Reliability Analysis

- Account for process variation

Energy Optimization

- Vary the application's schedule
- Minimize the system's energy
- Satisfy a number of constraints

Energy Optimization

Energy

$$\min_{\mathcal{S}} \mathbb{E}\left(E(\mathcal{S})\right)$$

$$\tau(\mathcal{S}) \leq \tau_{\mathrm{max}}$$

$$\mathbb{P}\left(Q(\mathcal{S}) \geq q_{\mathrm{max}}\right) \leq \rho_{\mathrm{burn}}$$

$$\mathbb{P}\left(\mathbb{E}\left(L(\mathcal{S})\right) \leq L_{\mathrm{min}}\right) \leq \rho_{\mathrm{wear}}$$

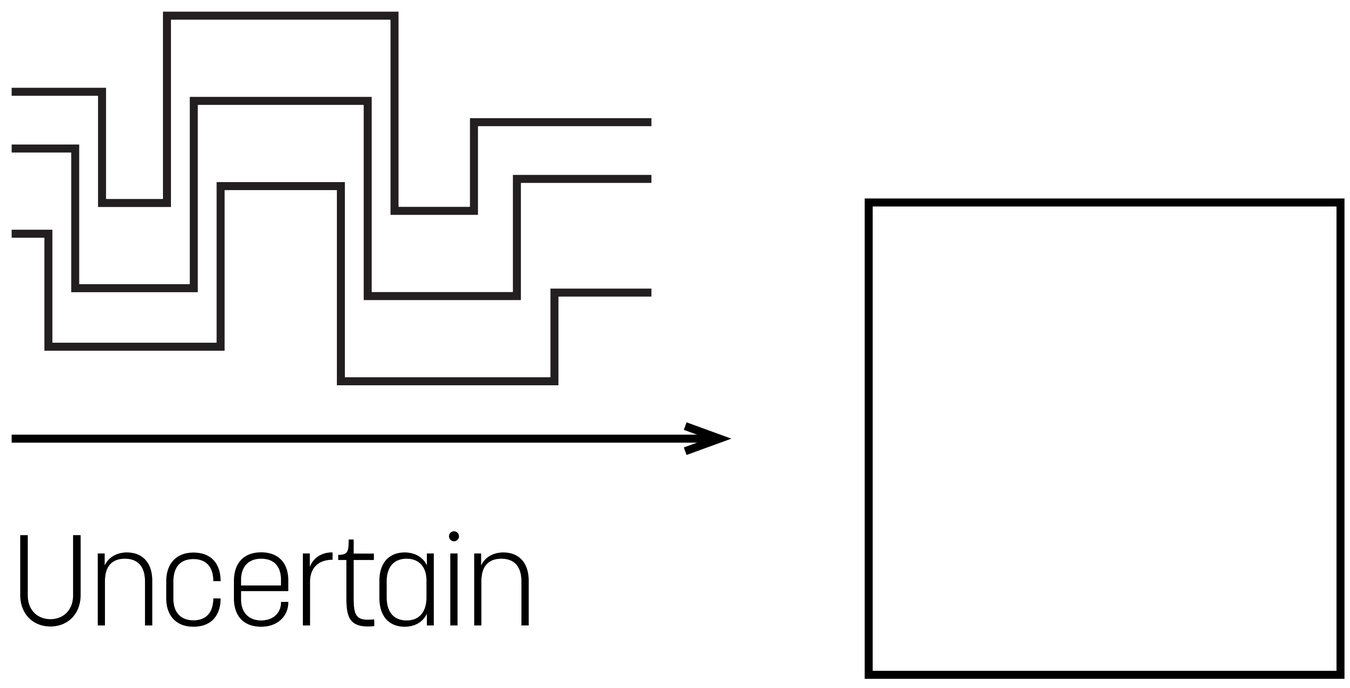
Lifetime

Experimental Results

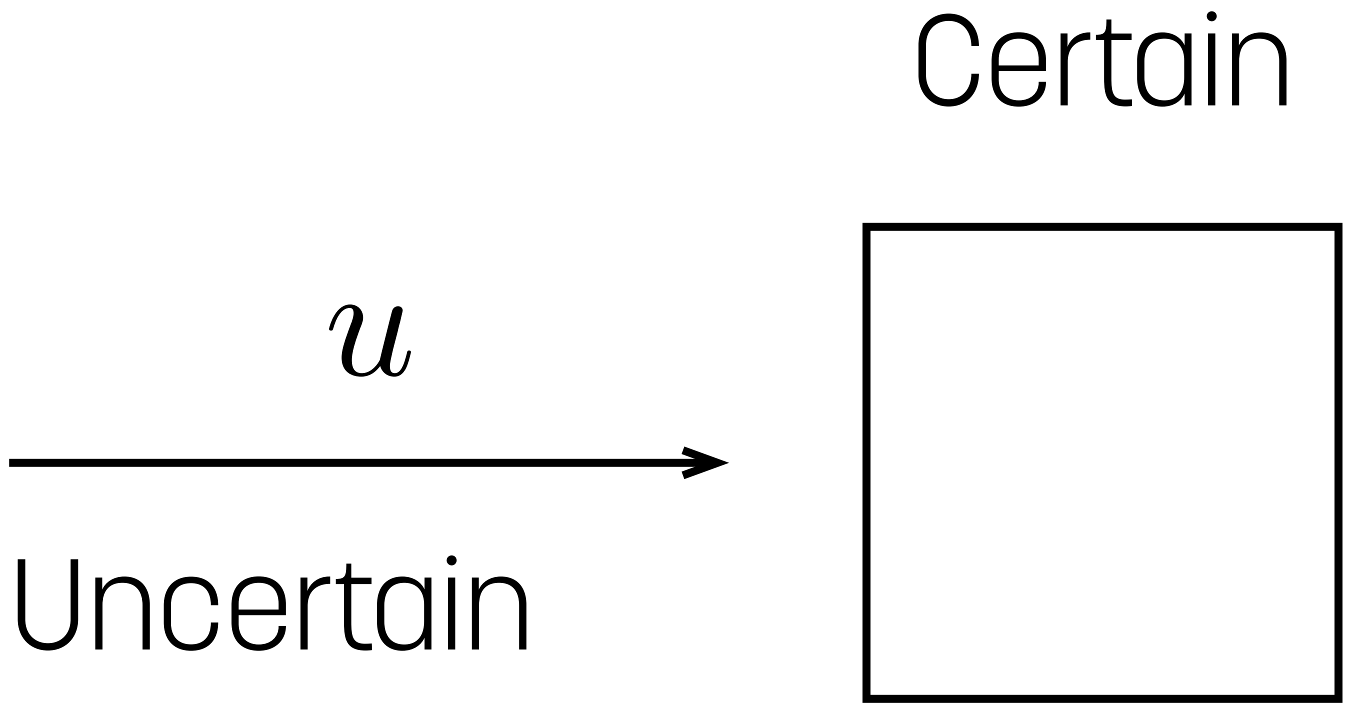
- Demonstrated the importance of accounting for process variation
    - Up to 100% of solutions that ignore uncertainty might be unacceptable

Analysis under Workload Variation

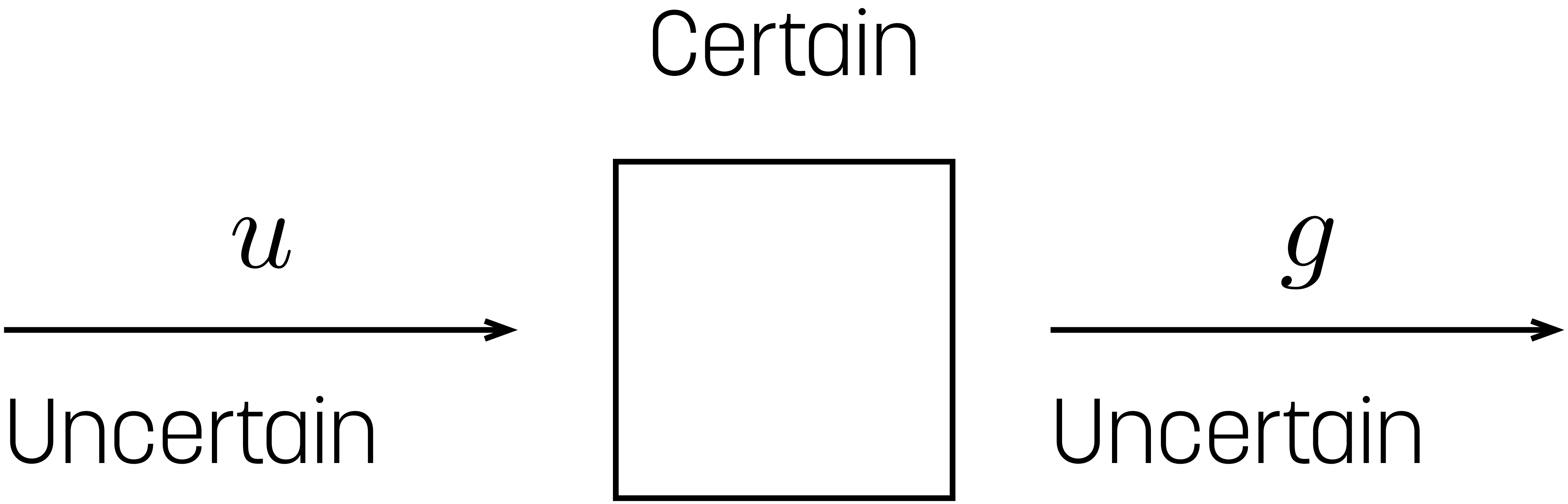
Workload Variation



Uncertain Parameters



Quantity of Interest

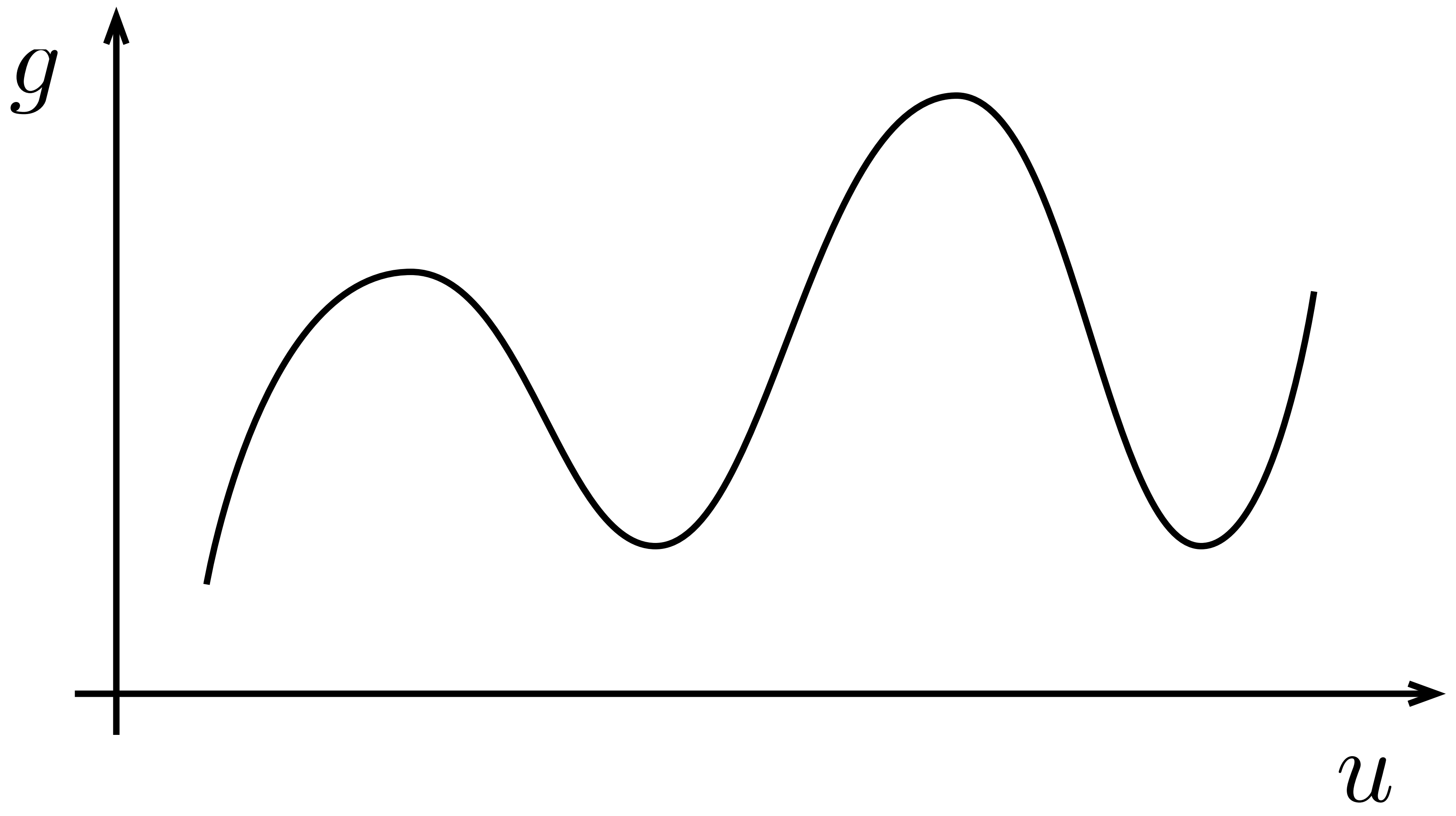


Problem Formulation

- Given the probability distribution of the uncertain parameters u
- Compute the probability distribution of the quantity of interest g

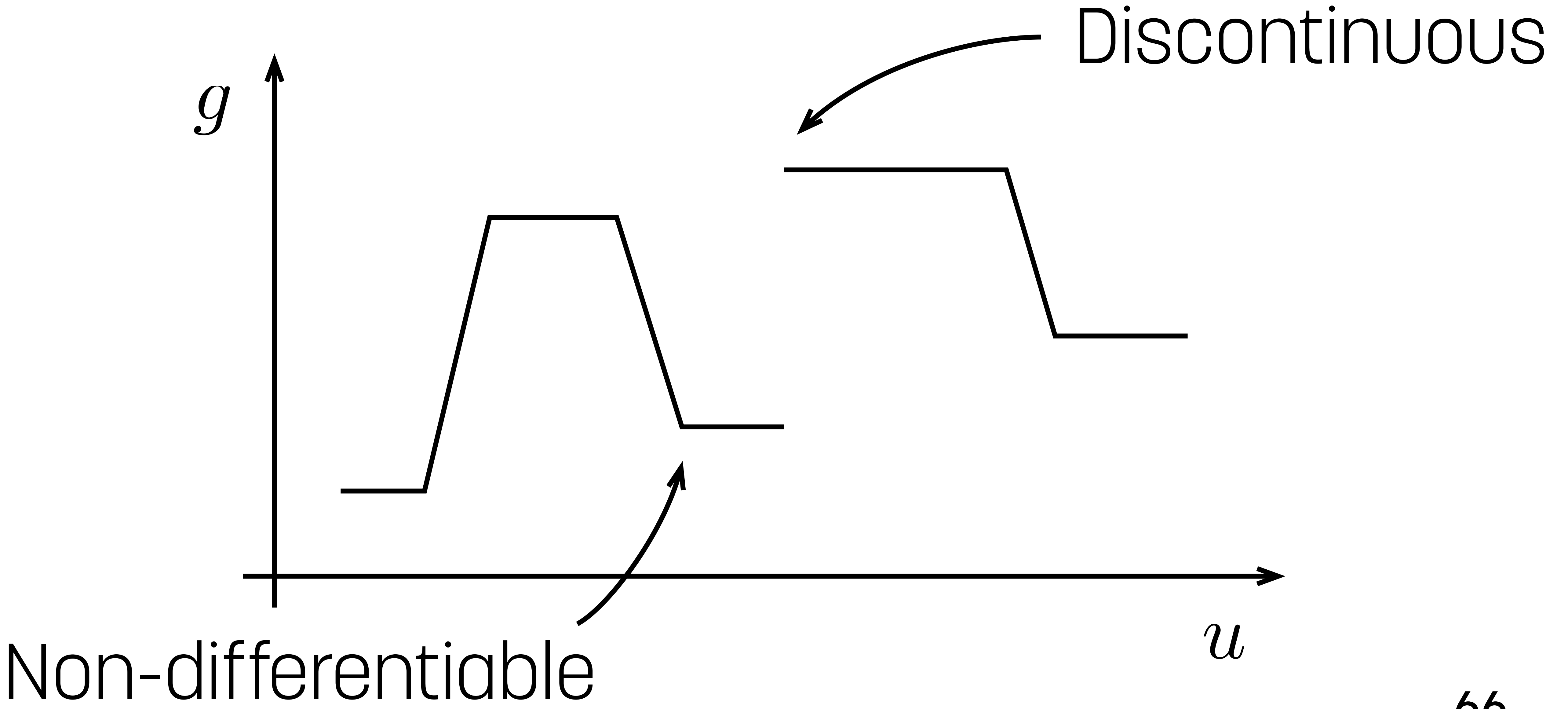
Response Surface

$$g = f(u)$$



Response Surface

$$g = f(u)$$



Response Surface

- Process variation
  - Smooth, well-behaved
- Workload variation
  - Non-smooth, ill-behaved

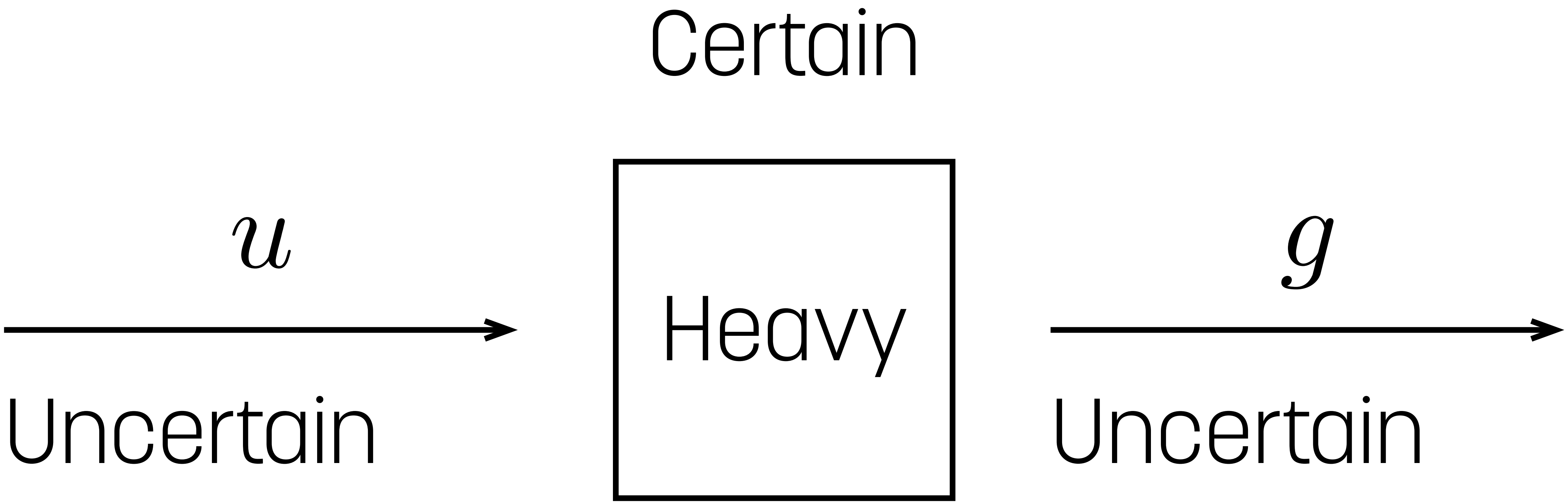
Previous Work

- Inadequate
- Limited in use

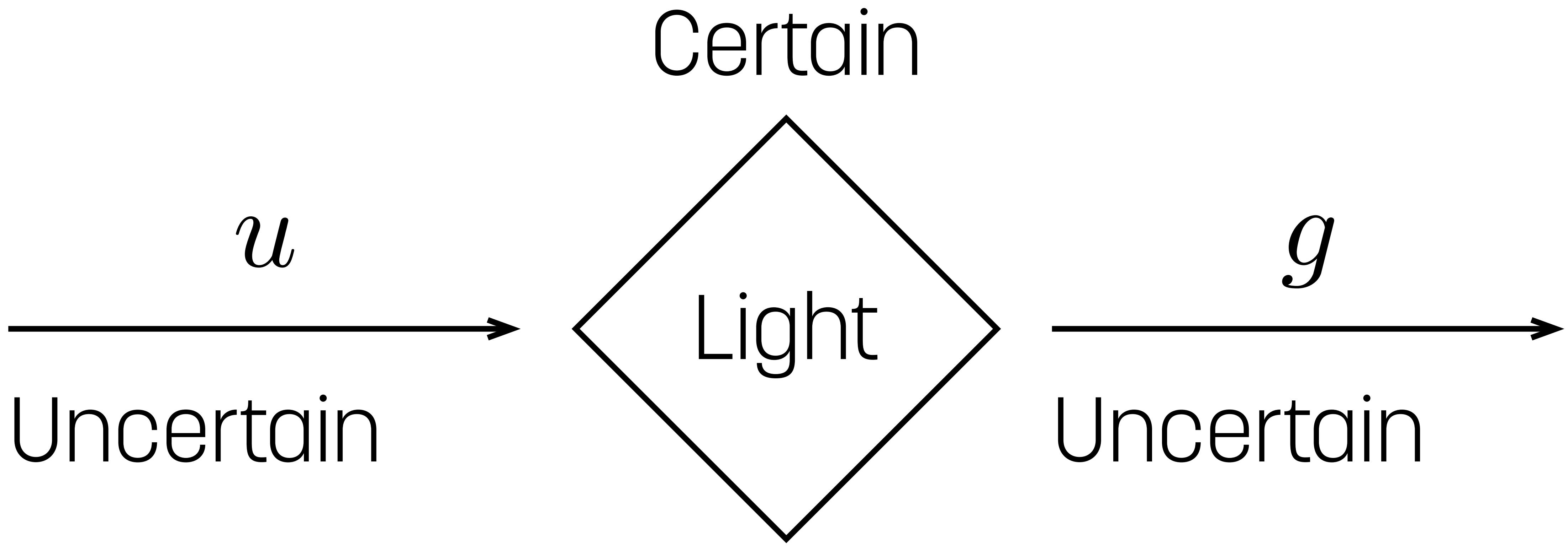
Proposed Solution

- General
- Efficient
- Easy to apply

Solution Overview



Solution Overview



Solution Overview

$$u = \mathbb{T}(z)$$

$$u : \Omega \to \mathbb{R}^{n_u}$$

$$z : \Omega \to [0,1]^{n_z}$$

$$g(u) = (g \circ \mathbb{T})(z) = g(\mathbb{T}(z))$$

Solution Overview

Adaptive hierarchical interpolation

$$g \approx \mathcal{A}_{l_s}^{n_z}(g) = \mathcal{A}_{l_s-1}^{n_z}(g) + \Delta\mathcal{A}_{l_s}^{n_z}(g)$$

Hierarchical surplus

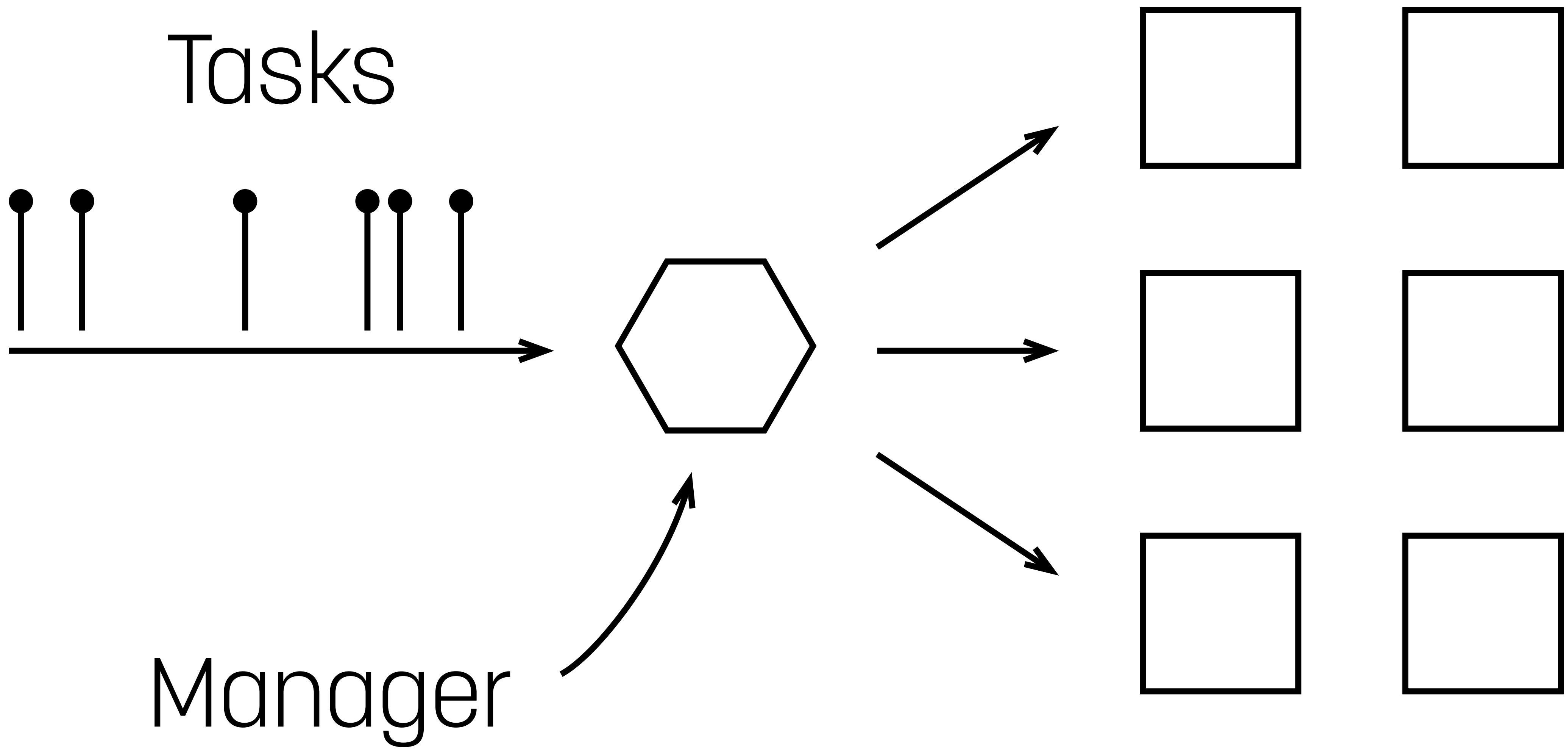
$$\Delta\mathcal{A}_{l_s}^{n_z}(g) = \sum_{i \in \Delta\mathcal{I}_{l_s}^{n_z}} \sum_{j \in \Delta\mathcal{J}_i^{n_z}} \Delta(g \circ \mathbb{T})(x_{ij}) e_{ij}$$

Experimental Results

- Applied to a set of synthetic problems and to a real-life problem
- Shown computational efficiency
  - 1–2 orders of magnitude more accurate than direct sampling

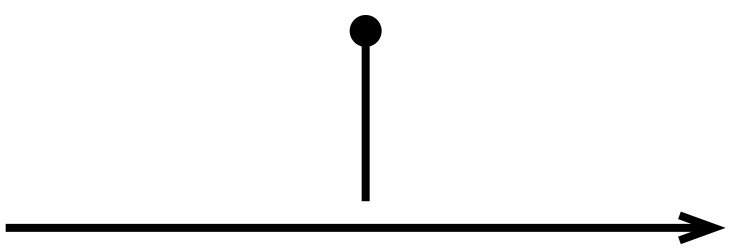
Resource Management under Workload Variation

Resource Management

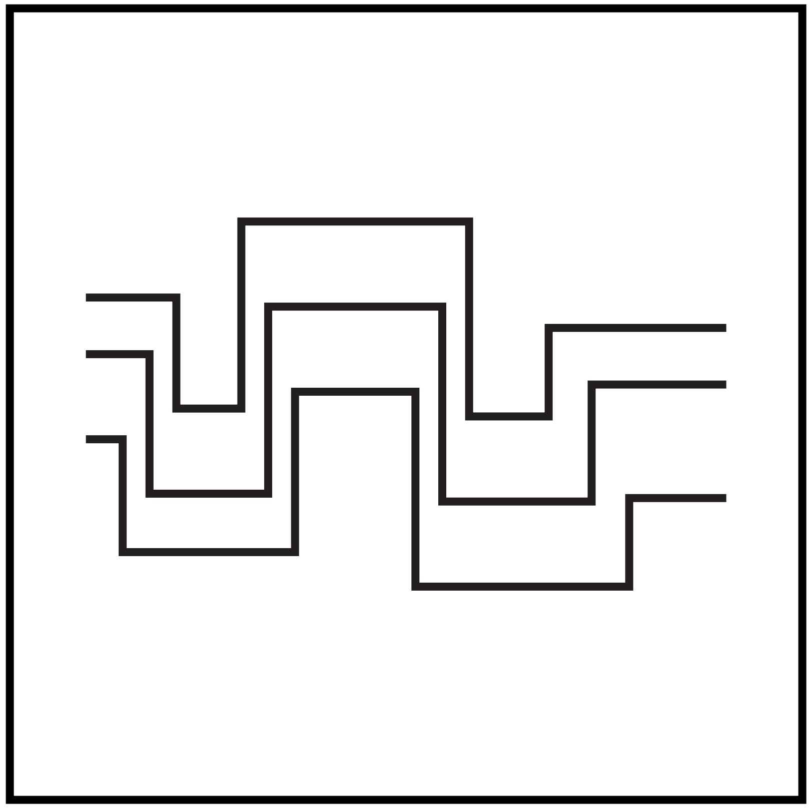


Resource Usage

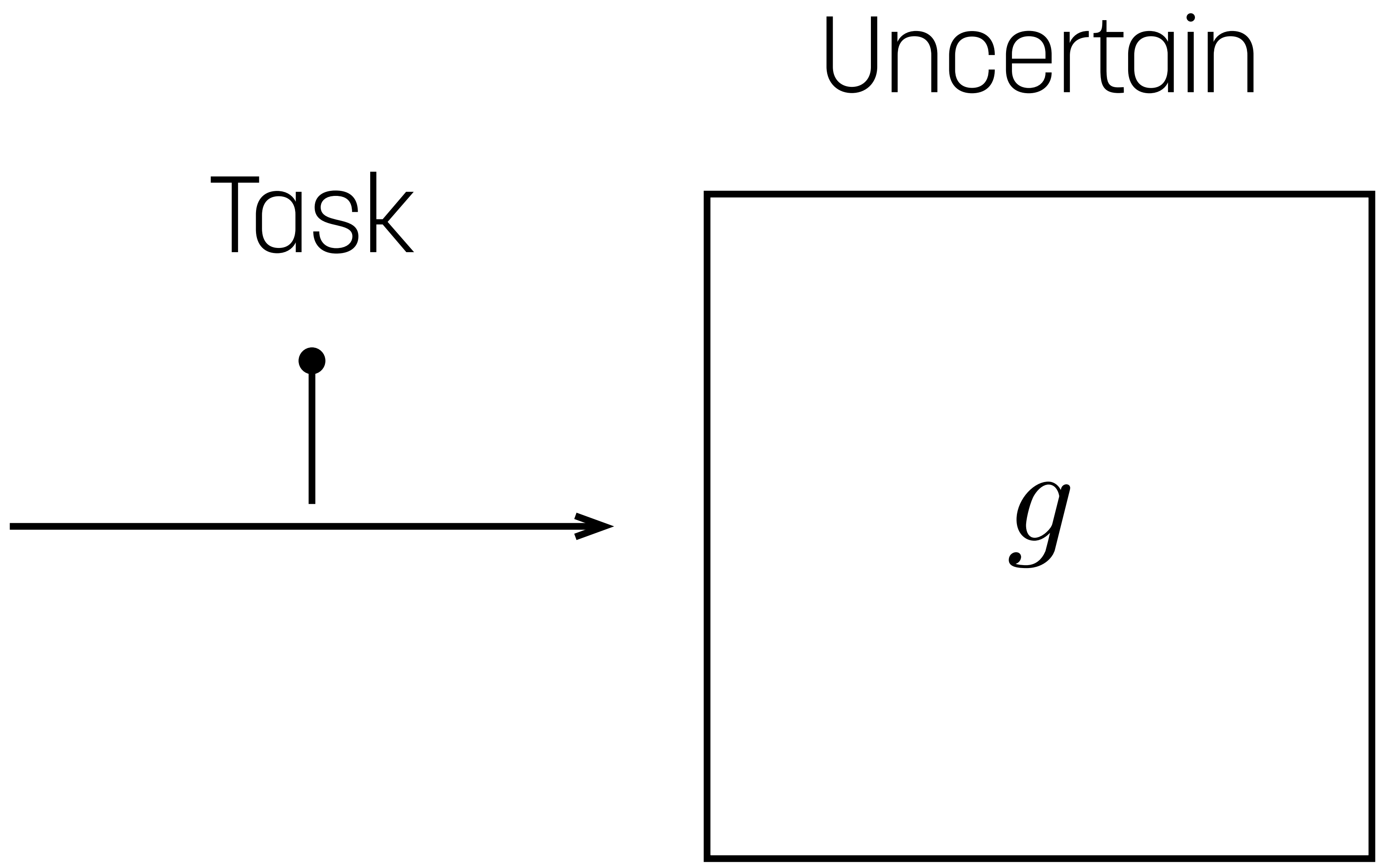
Task



Uncertain



Quantity of Interest



Problem Formulation

- Given past resource-usage traces G
- Predict resource usage for individual tasks multiple steps ahead

Previous Work

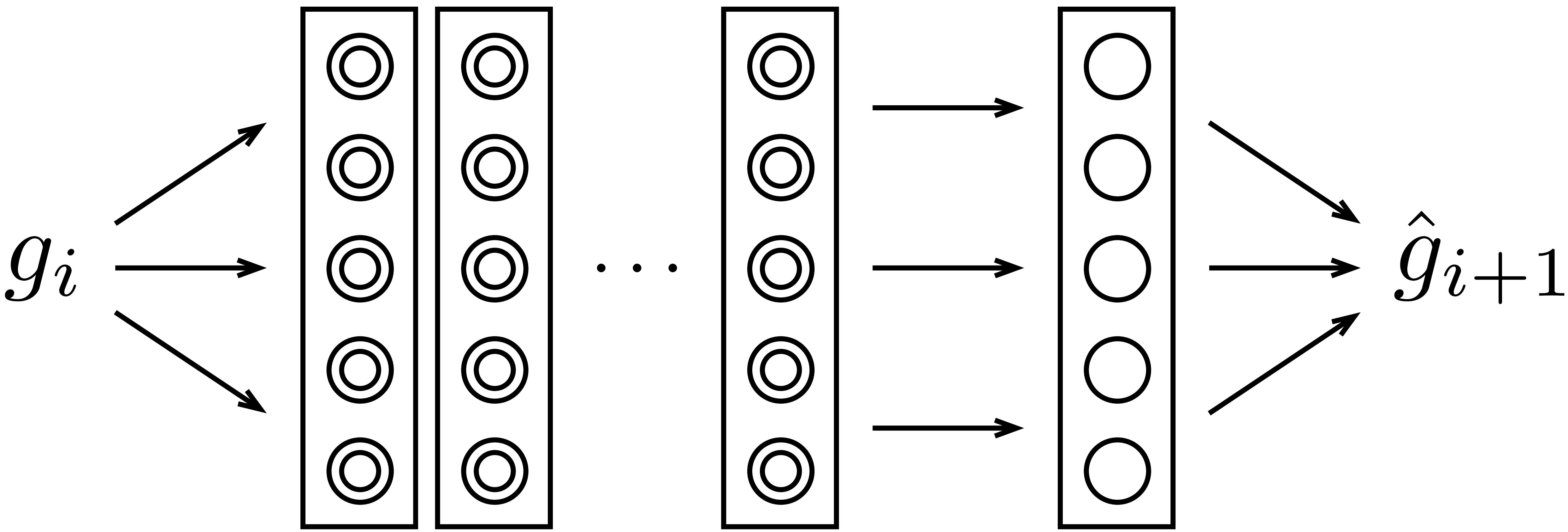
- Nonexistent
    - Only aggregate

Proposed Solution

- Fine grained
- Long range

Solution Overview

Recurrent neural network



Experimental Results

- Studied CPU usage in Google's computer cluster with 12500 nodes
- Shown the existence of a structure suitable for educated prediction
    - Error reduction of 47% for 4 steps ahead compared to random walk

Conclusion

Outline

1. Analysis and Design with Certainty
2. Characterization of Process Variation
3. Analysis and Design under Process Variation
4. Analysis under Workload Variation
5. Resource Management under Workload Variation

Open Source

- https://github.com/learning-on-chip
- https://github.com/markov-chain
- https://github.com/math-rocks
- https://github.com/ready-steady
- https://github.com/stainless-steel
- https://github.com/turing-complete