Energy-Efficient Mobile Web Computing

Yuhao Zhu
Electrical and Computer Engineering Department
The University of Texas at Austin
Advisor: Vijay Janapa Reddi

The University of Texas at Austin

Sept. 7th, 2016

The Web Evolution

1990

HTML

The Web Evolution

1990

HTML

The Web Evolution

1990

HTML

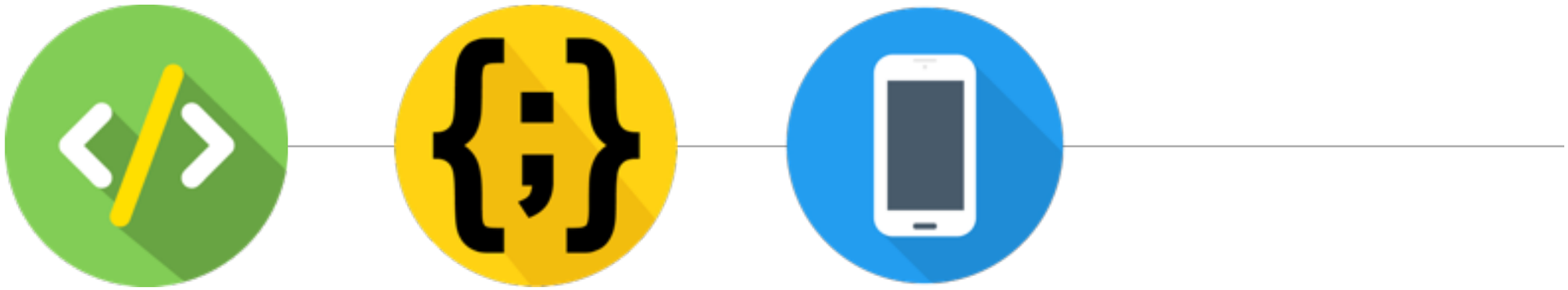
1996

JavaScript

The Web Evolution

1990
HTML

2008
Mobile Web



1996
JavaScript

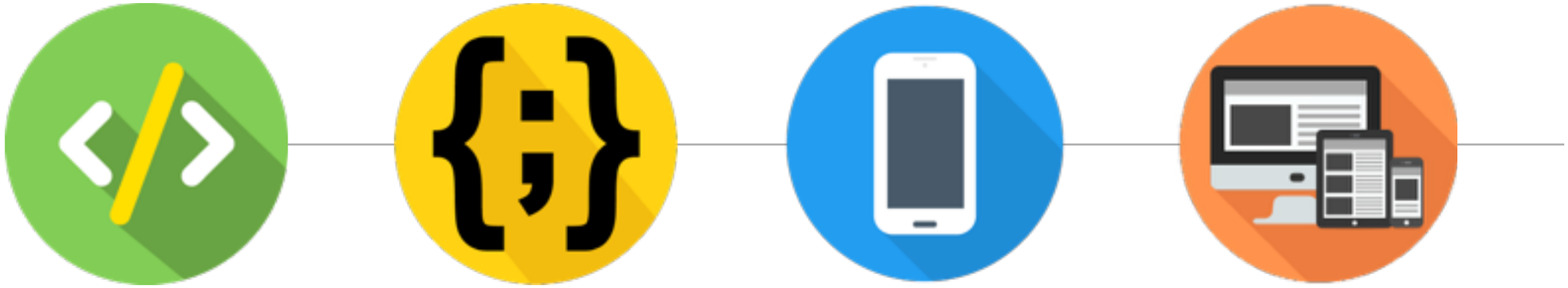
The Web Evolution

1990
HTML

1996
JavaScript

2008
Mobile Web

2012
Responsive Web

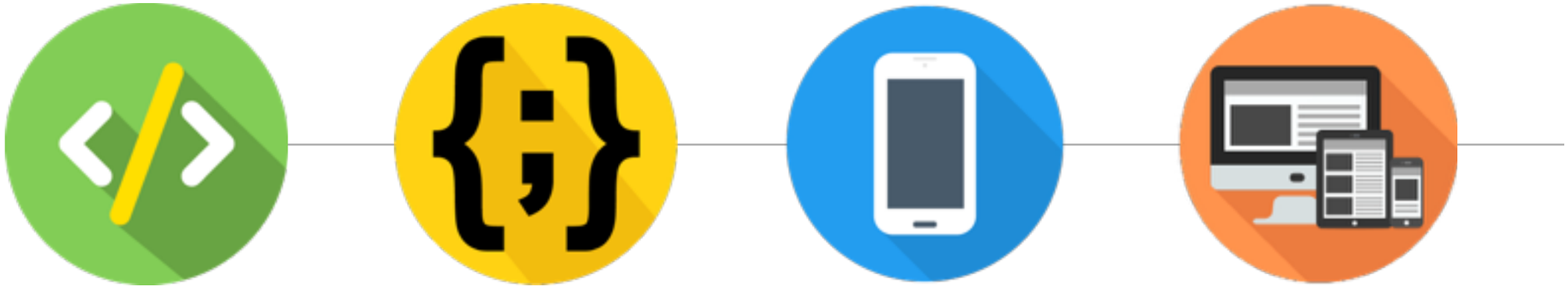


The Web Evolution

Functionality

1990
HTML

2008
Mobile Web



1996
JavaScript

2012
Responsive
Web

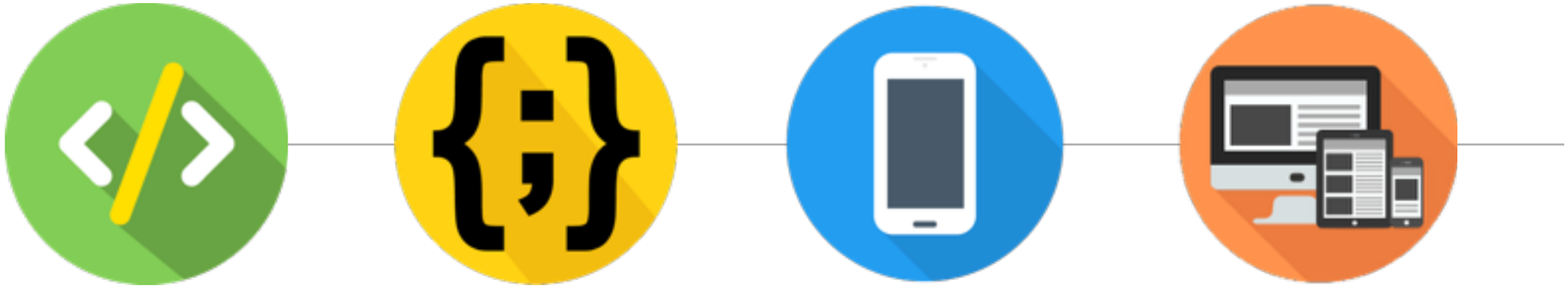
The Web Evolution

Functionality

Performance

1990
HTML

2008
Mobile Web



1996
JavaScript

2012
Responsive
Web

The Web Evolution

Functionality → Performance → Energy

1990
HTML

1996
JavaScript

2008
Mobile Web

2012
Responsive Web

2016
Watt Wise Web

The Web Evolution

Functionality

Performance

Energy

1990
HTML

1996
JavaScript

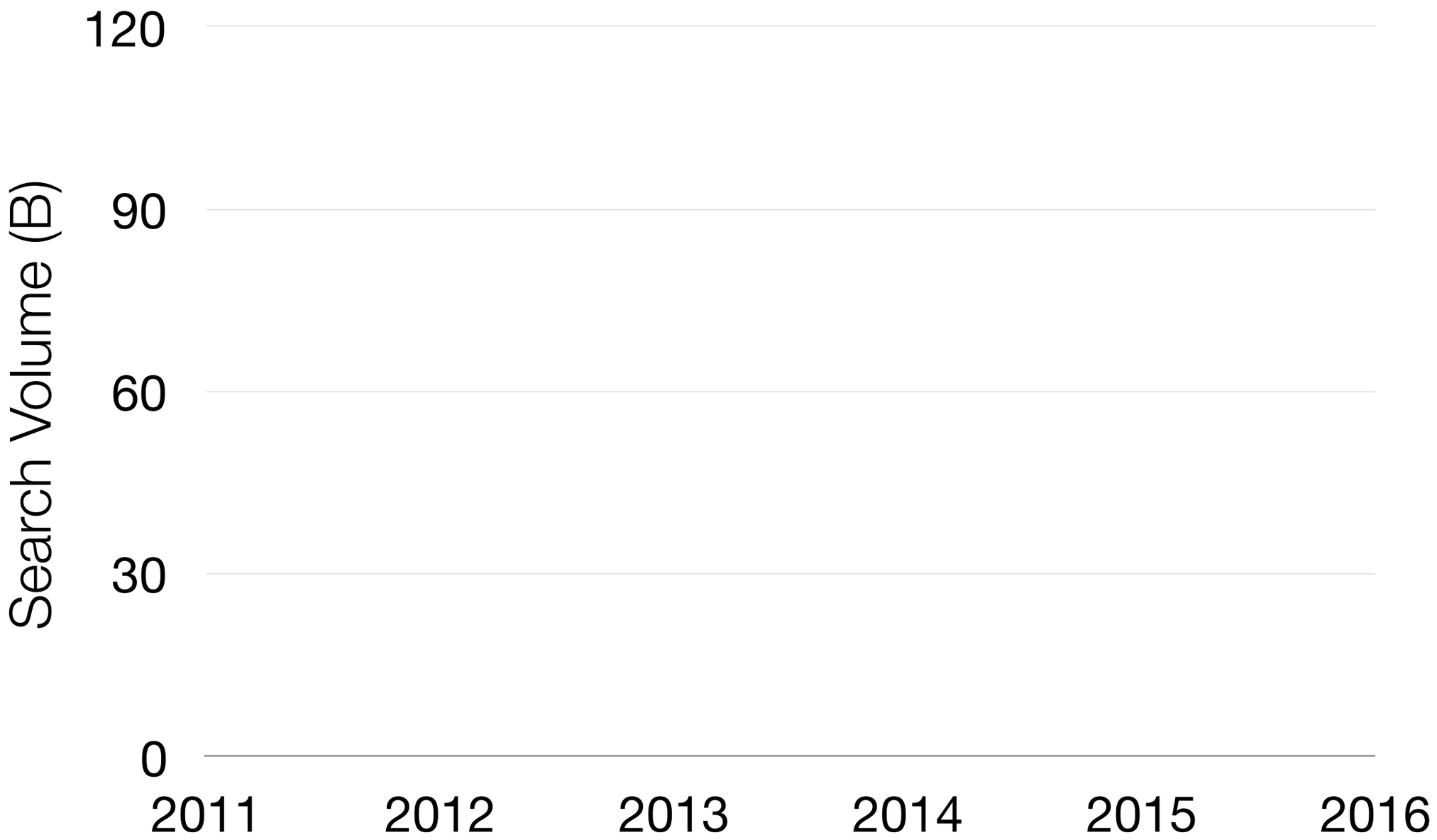
2008
Mobile Web

2012
Responsive Web

2016
Watt Wise Web

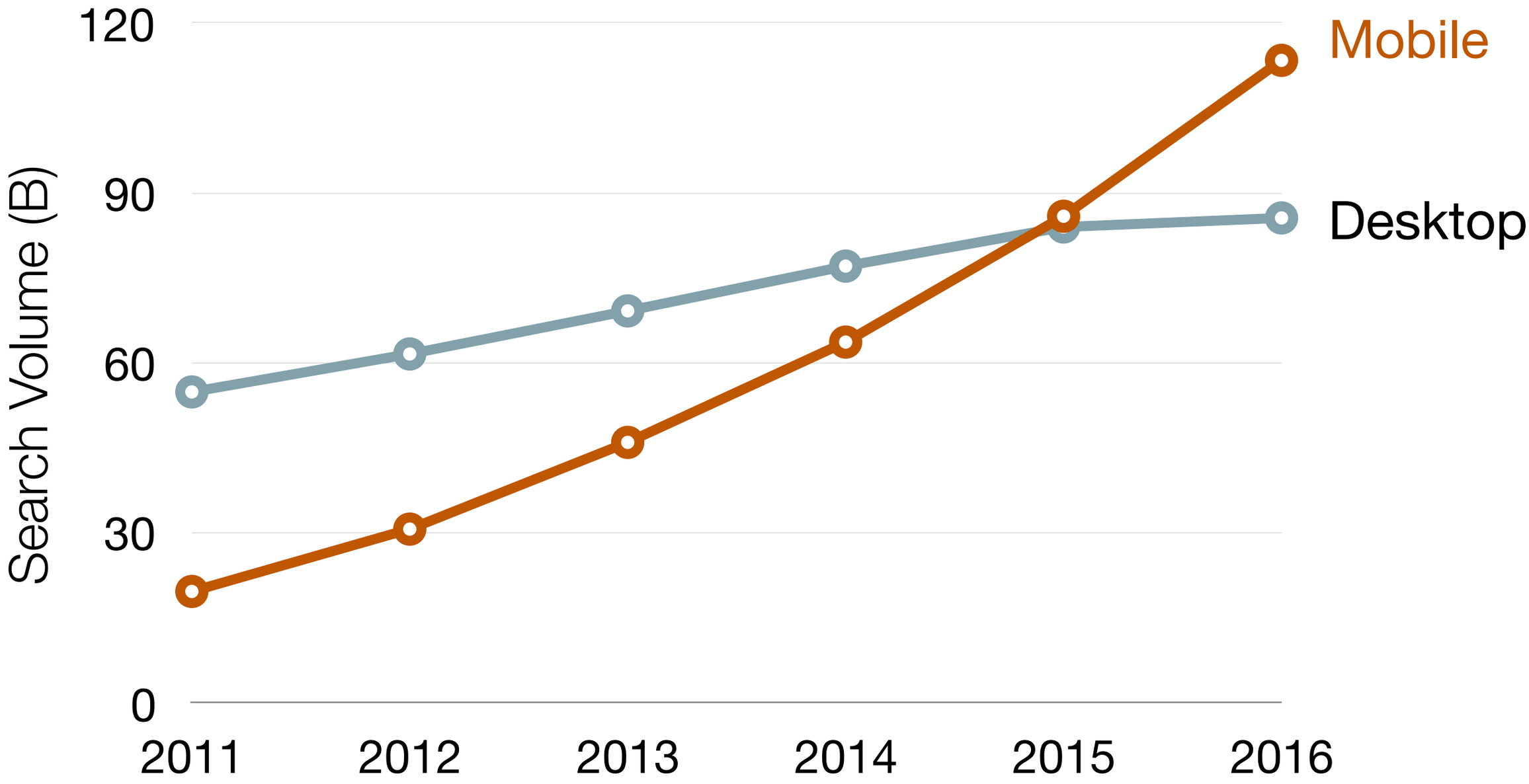
Web: Mobile Overtaking Desktop

Web: Mobile Overtaking Desktop



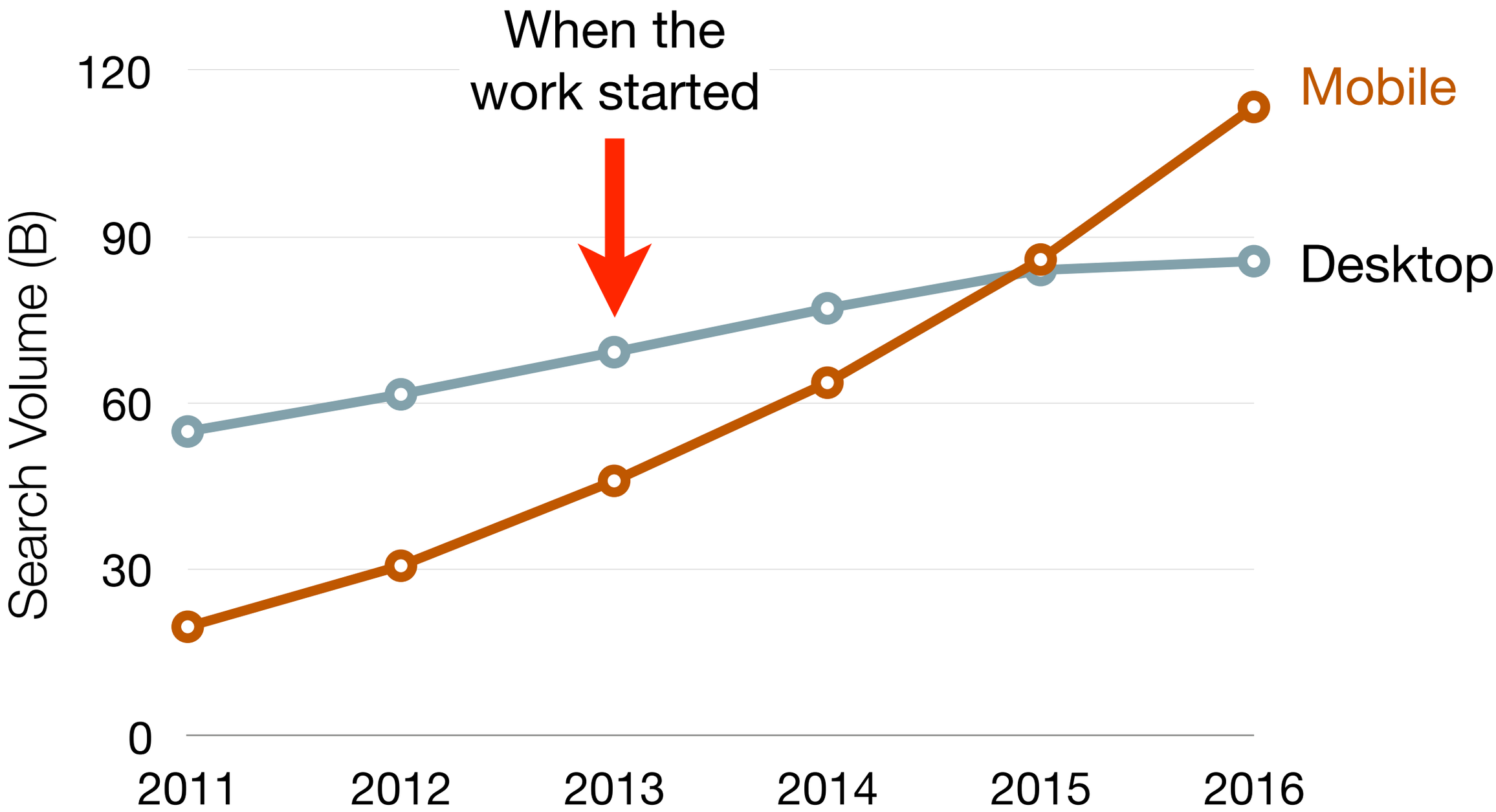
Source: BIA/Kelsey

Web: Mobile Overtaking Desktop



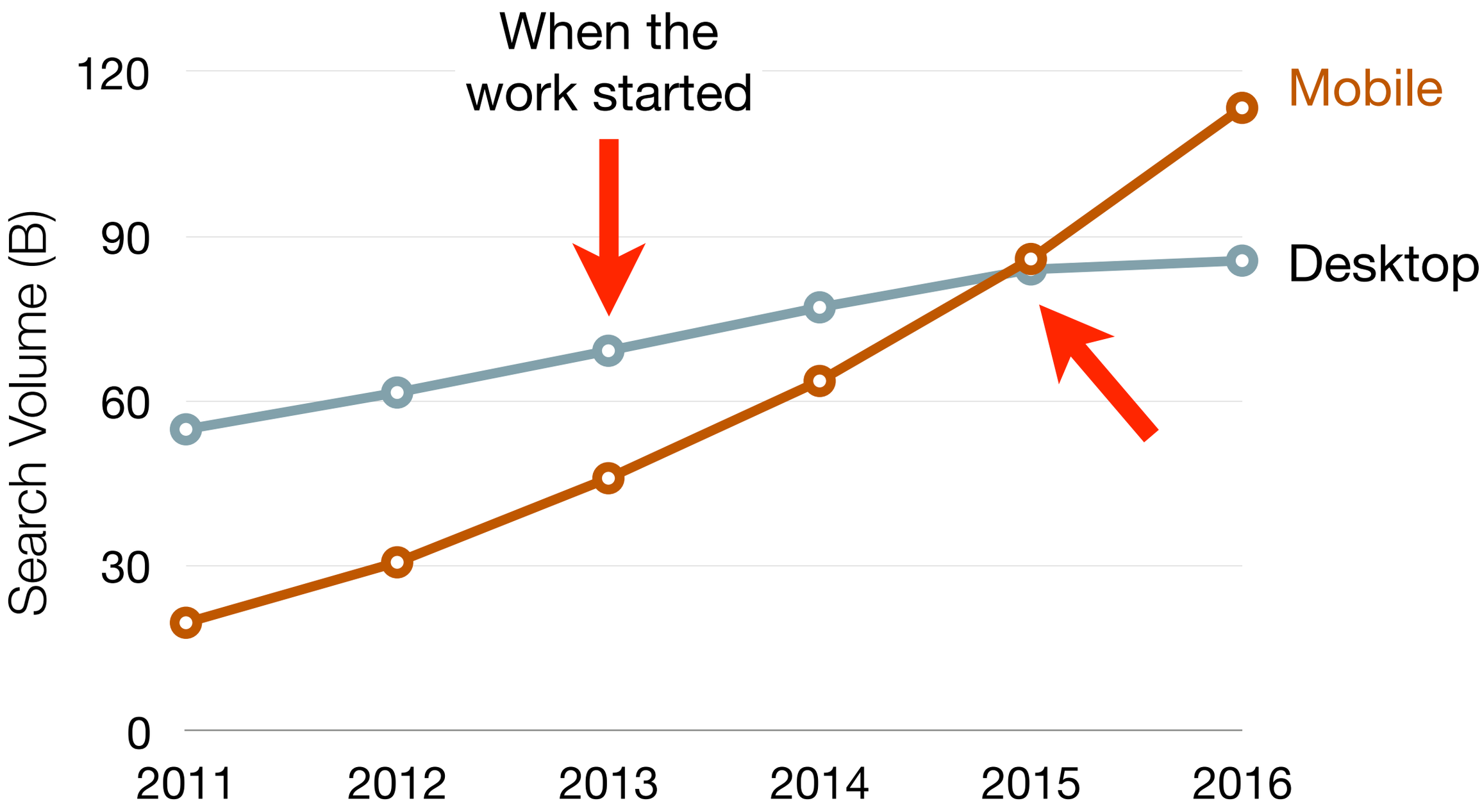
Source: BIA/Kelsey

Web: Mobile Overtaking Desktop



Source: BIA/Kelsey

Web: Mobile Overtaking Desktop



Source: BIA/Kelsey

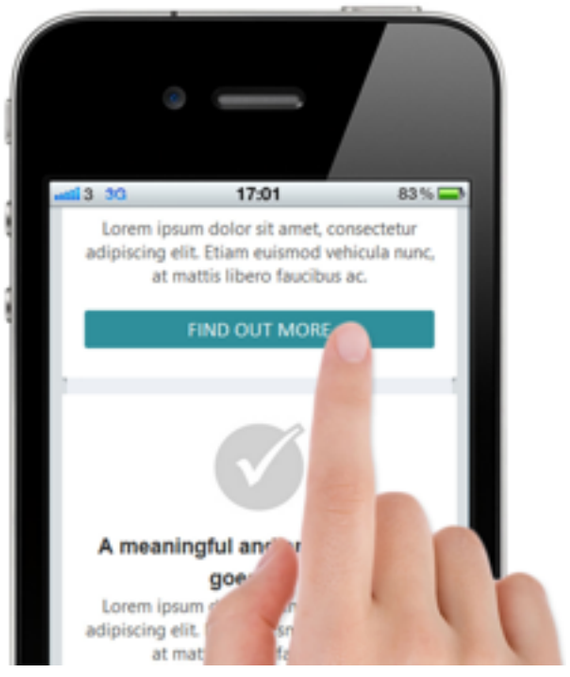
Web ≈ Mobile Web

The Scope of Mobile Web

Mobile
Client

The Scope of Mobile Web

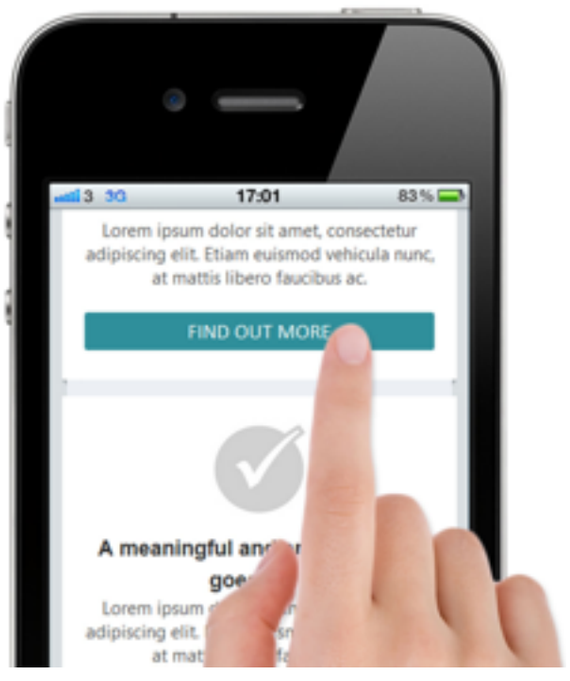
Mobile Client



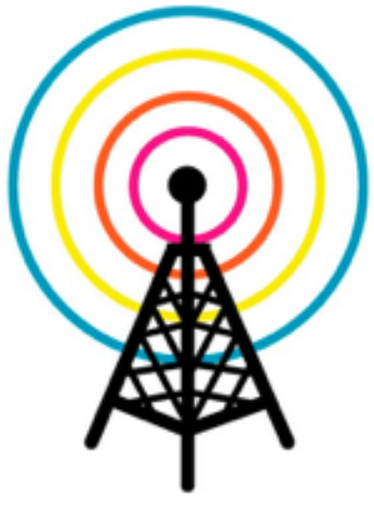
Cloud Web Servers

The Scope of Mobile Web

Mobile Client



Cellular Network



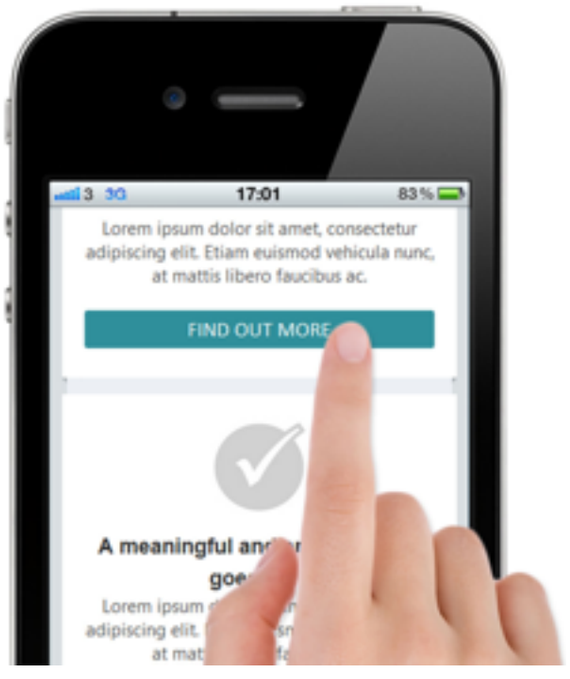
Cloud Web Servers

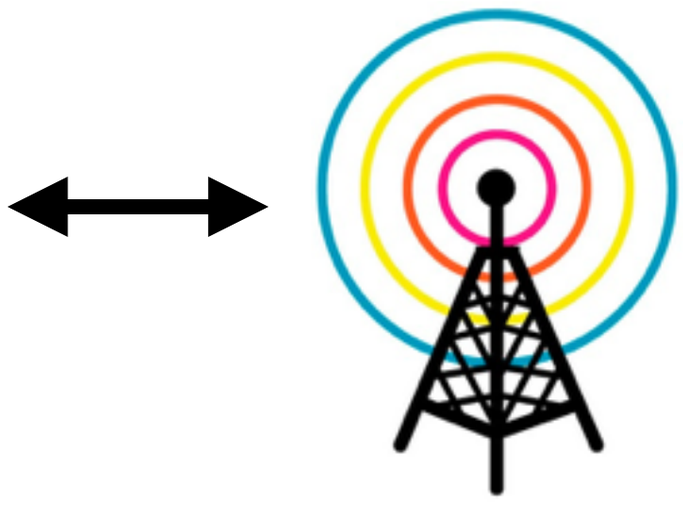
The Scope of Mobile Web

Mobile Client

Cellular Network

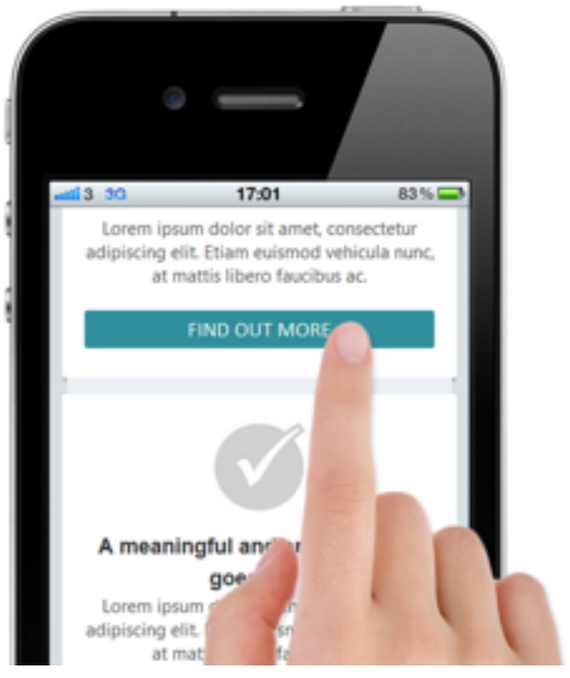
Cloud Web Servers



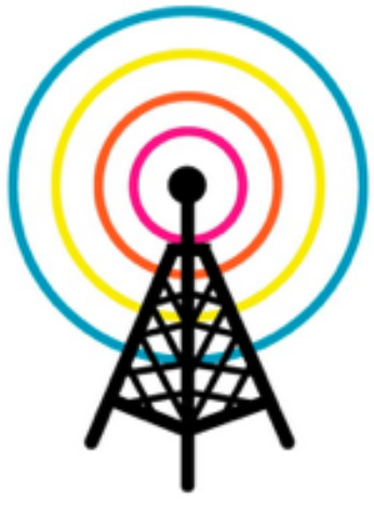


The Scope of Mobile Web

Mobile Client



Cellular Network



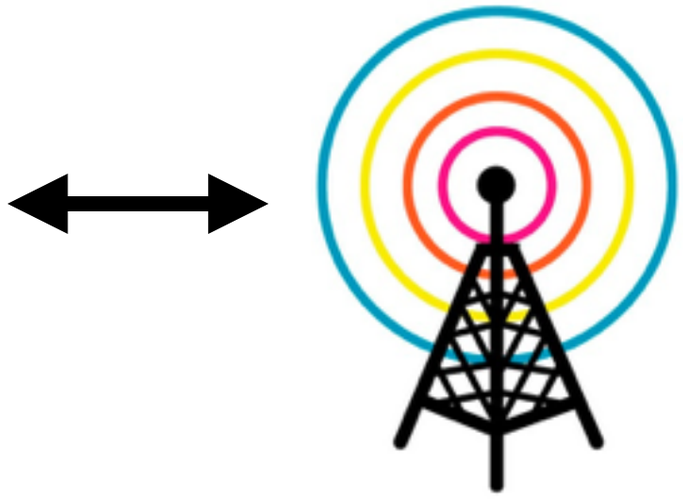
Cloud Web Servers

[MICRO 2015] (Top Picks Honorable Mention)

The Scope of Mobile Web

Mobile Client

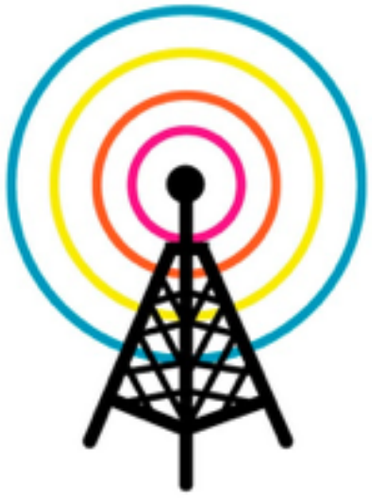
Cellular Network



The Scope of Mobile Web

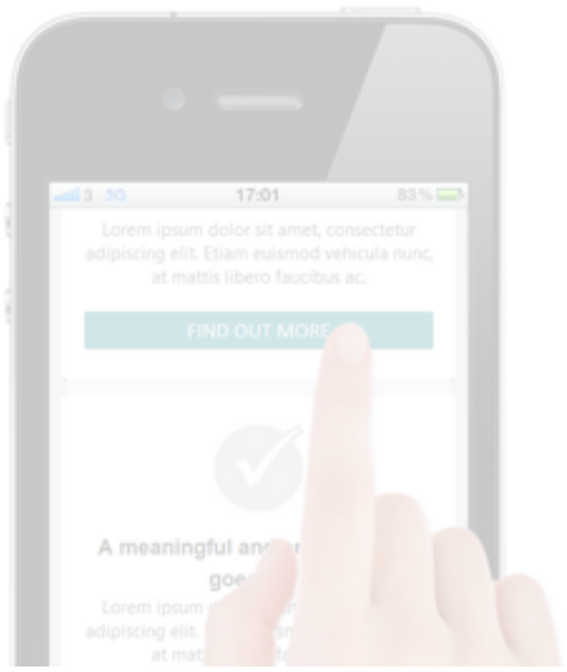
Mobile Client

Cellular Network

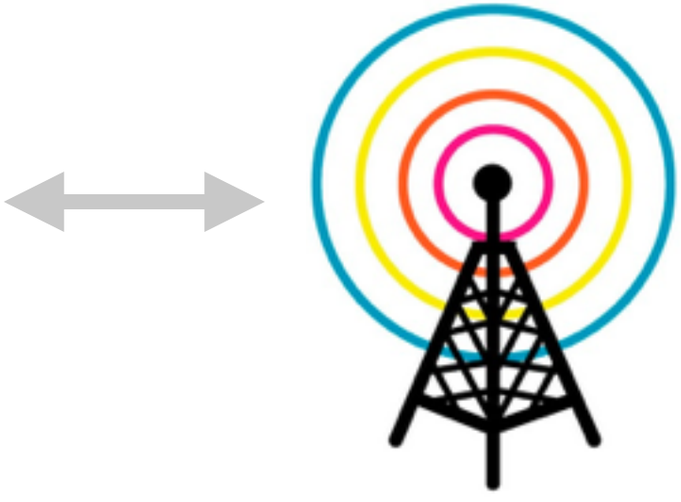


Isn’t Mobile Web a Network Issue?

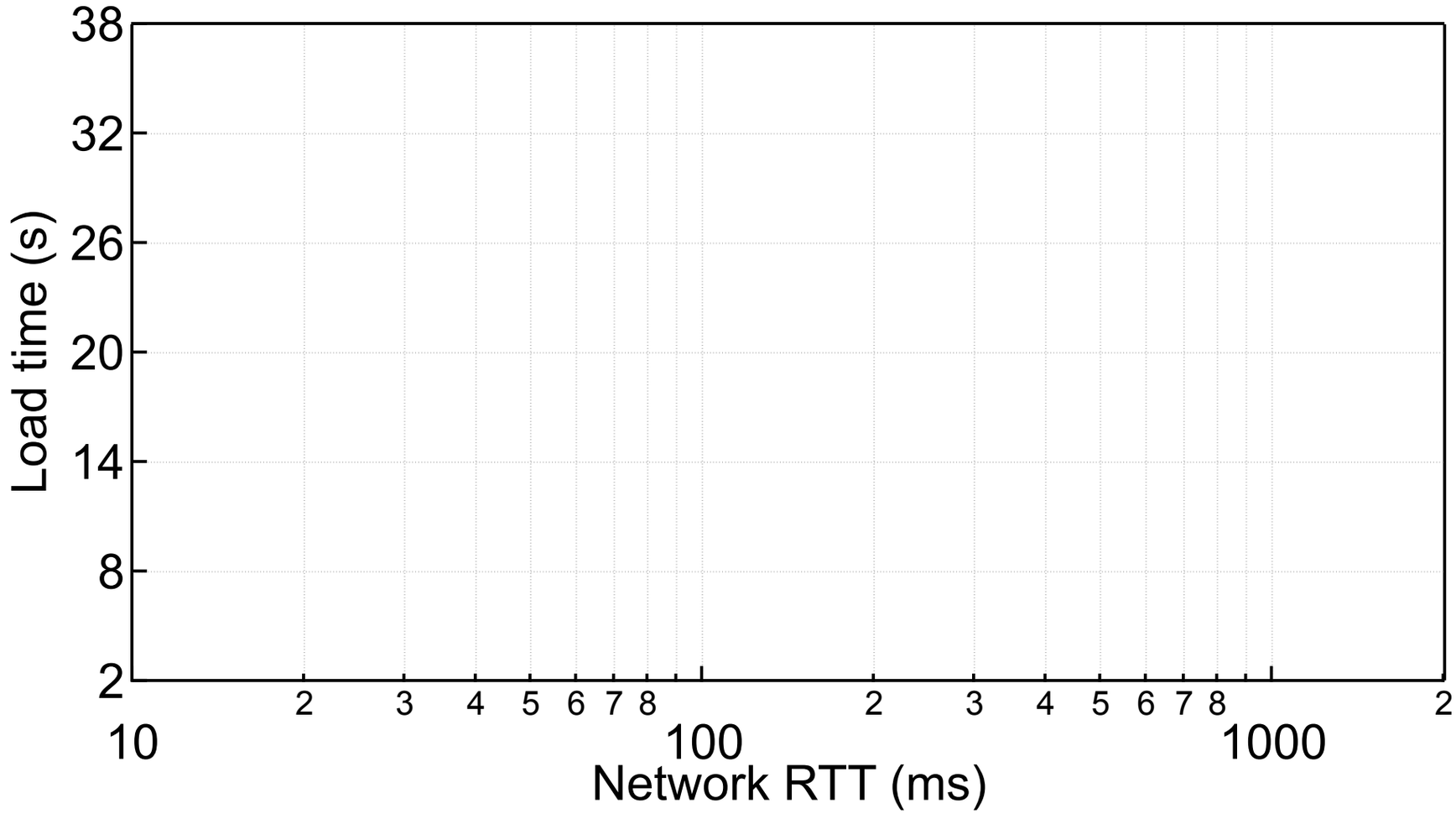
Mobile Client



Cellular Network

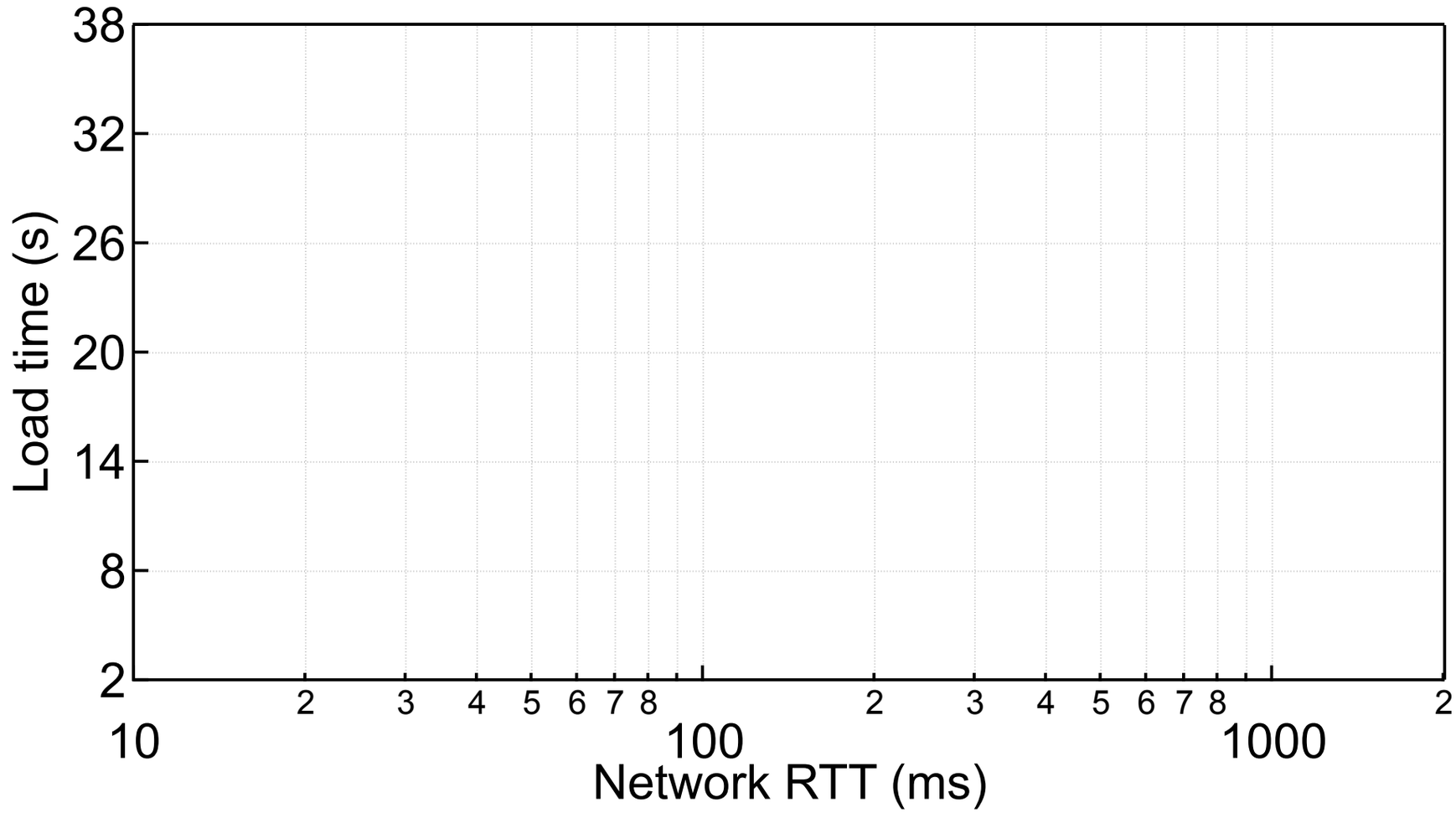


Isn’t Mobile Web a Network Issue?



Isn’t Mobile Web a Network Issue?

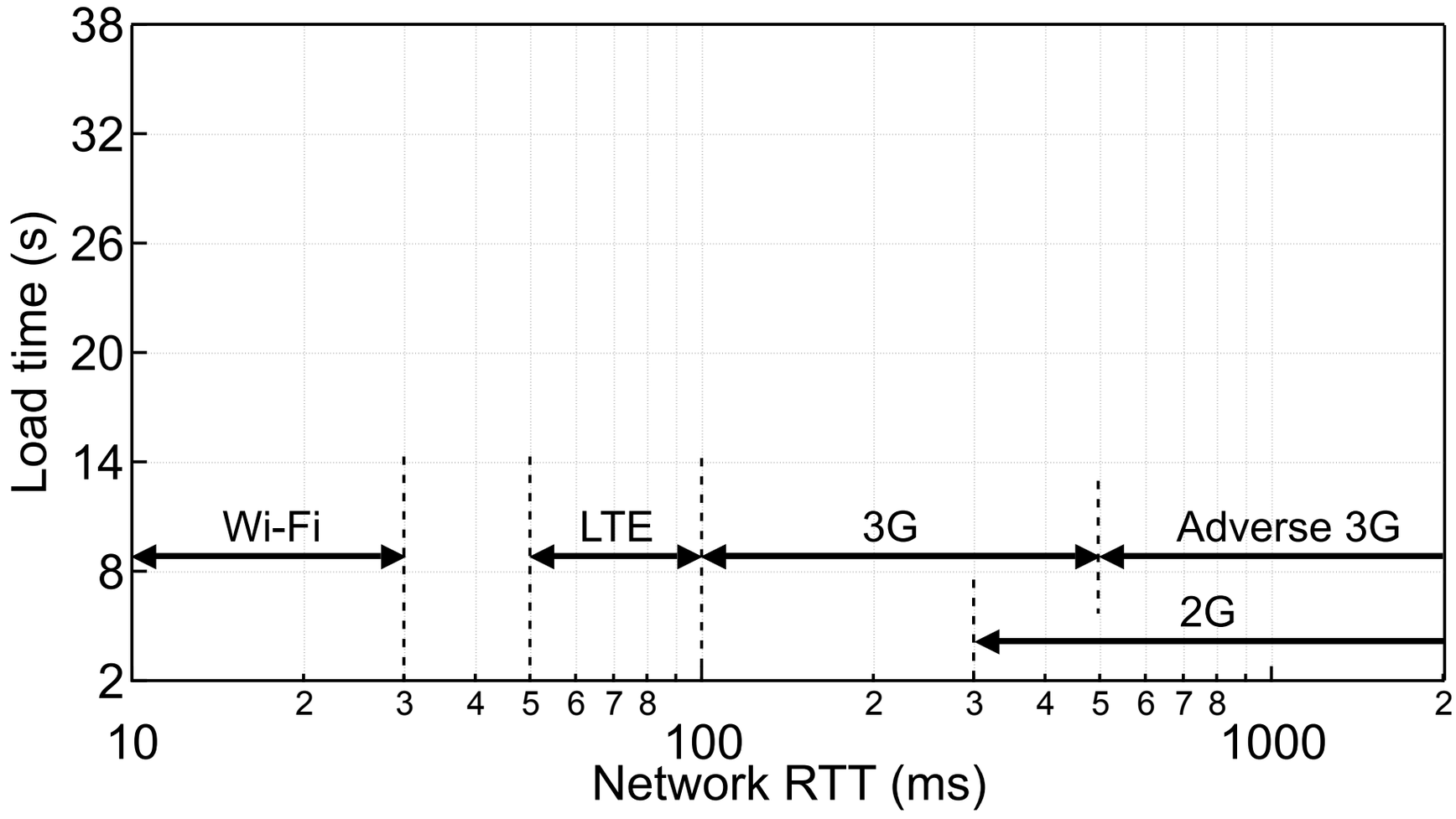
- Samsung Galaxy S4 smartphone.
- Hot webpages from Alexa[1].
- Time measured using Navigation Timing API[2].



1. http://www.alexa.com/
2. https://www.w3.org/TR/navigation-timing-2/

Isn’t Mobile Web a Network Issue?

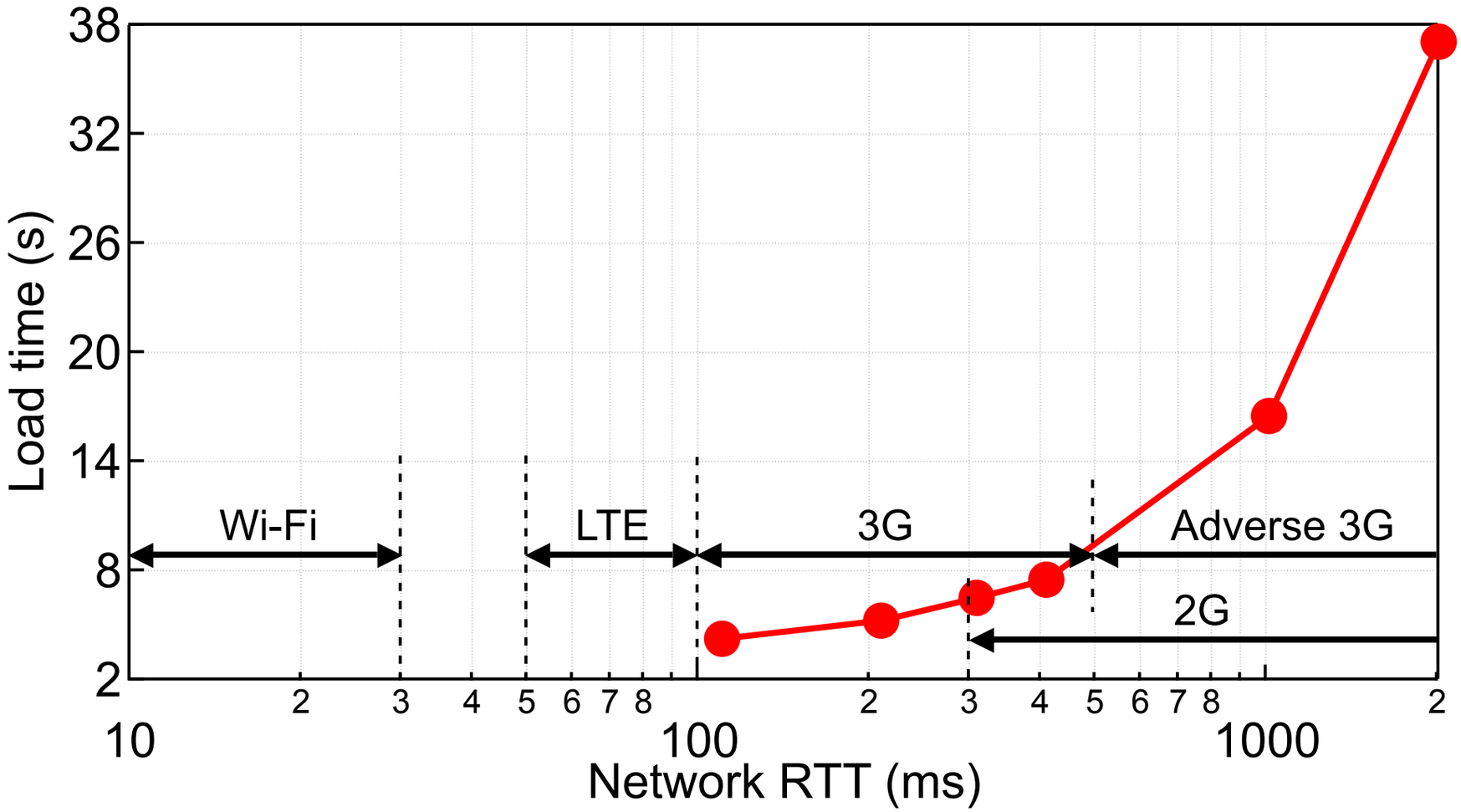
- Samsung Galaxy S4 smartphone.
- Hot webpages from Alexa[1].
- Time measured using Navigation Timing API[2].



1. http://www.alexa.com/
2. https://www.w3.org/TR/navigation-timing-2/

Isn’t Mobile Web a Network Issue?

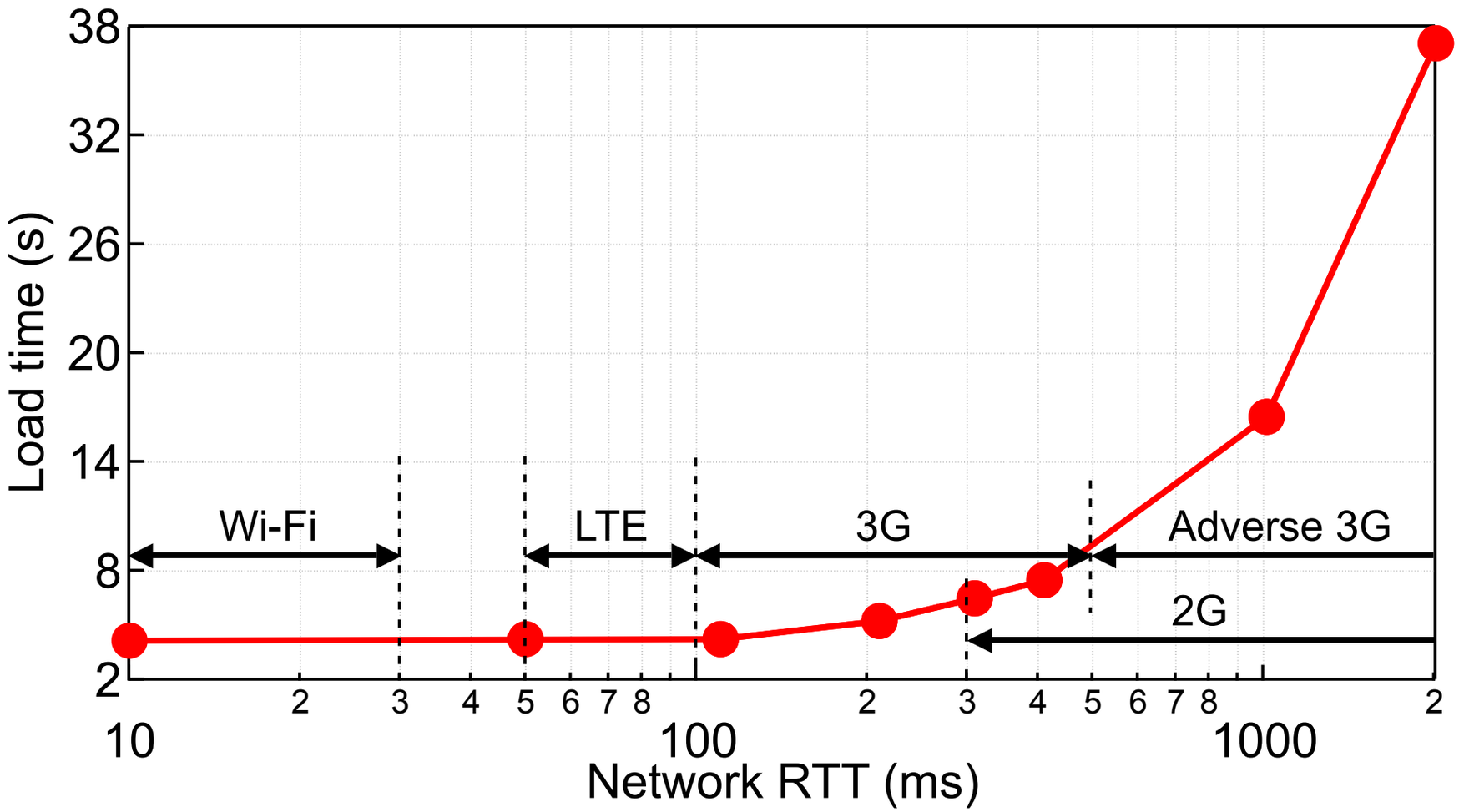
- Samsung Galaxy S4 smartphone.
- Hot webpages from Alexa[1].
- Time measured using Navigation Timing API[2].



1. http://www.alexa.com/
2. https://www.w3.org/TR/navigation-timing-2/

Isn’t Mobile Web a Network Issue?

- Samsung Galaxy S4 smartphone.
- Hot webpages from Alexa[1].
- Time measured using Navigation Timing API[2].

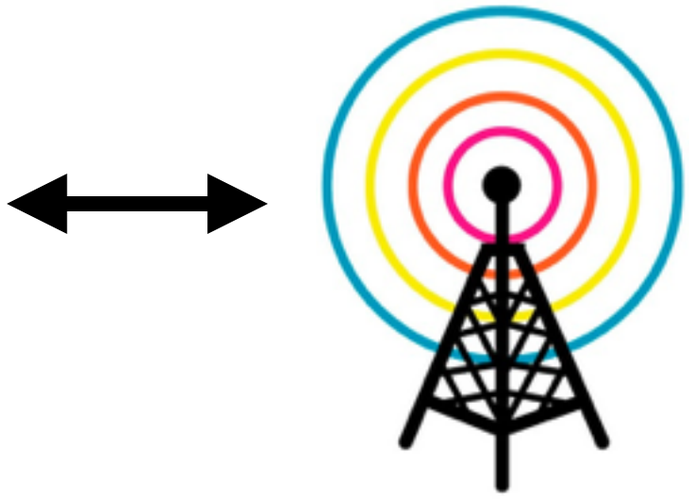


1. http://www.alexa.com/
2. https://www.w3.org/TR/navigation-timing-2/

Mobile Web is also a Compute Issue!

Mobile Client

Cellular Network



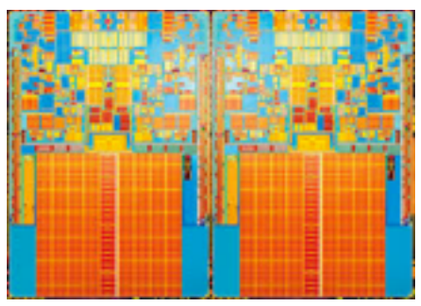
Mobile Web is also a Compute Issue!

Mobile Client

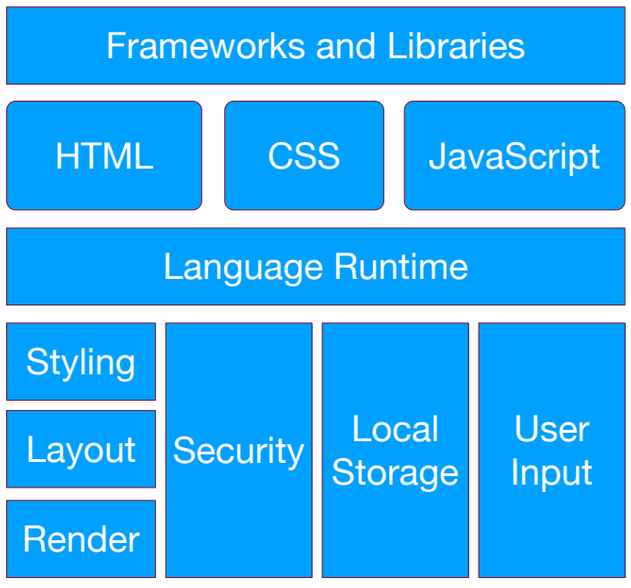
Cellular Network

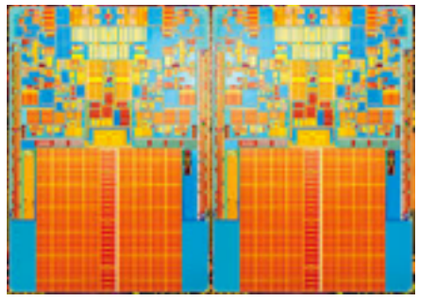
My Work

Traditional Approach

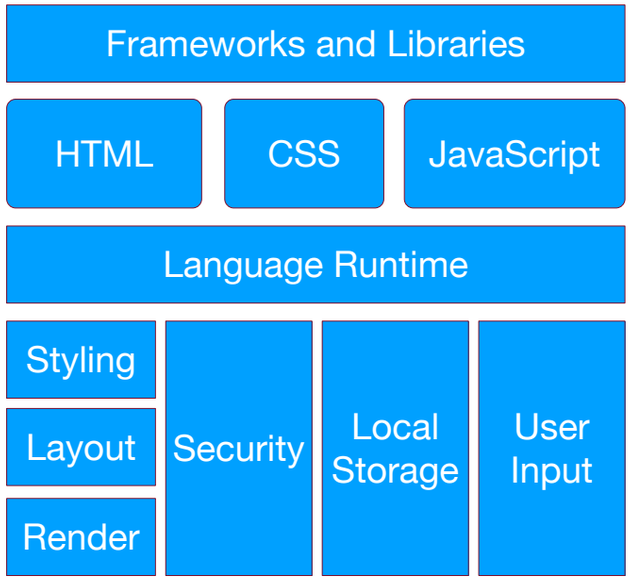


Traditional Approach

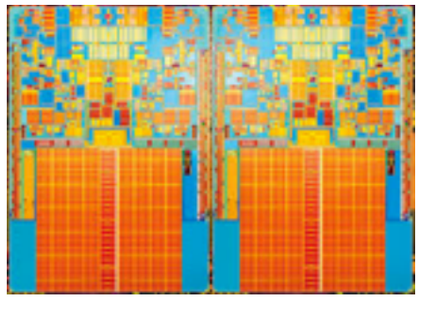




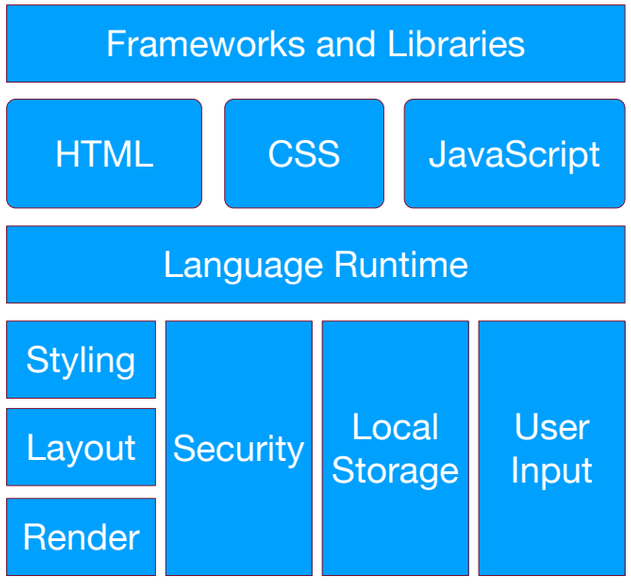
Traditional Approach

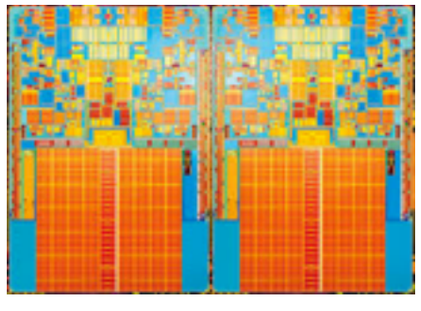


Application



Traditional Approach

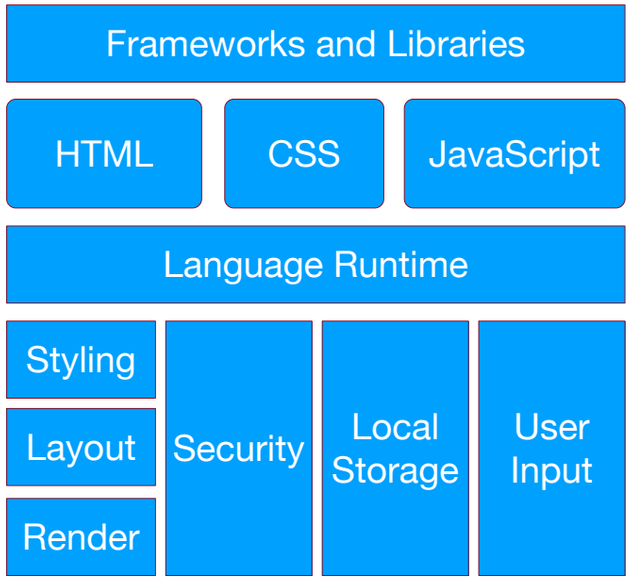


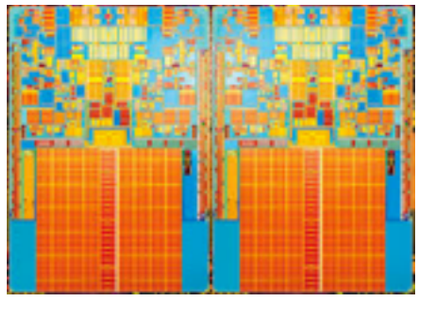


Application

- Parallelize browser computation

Traditional Approach



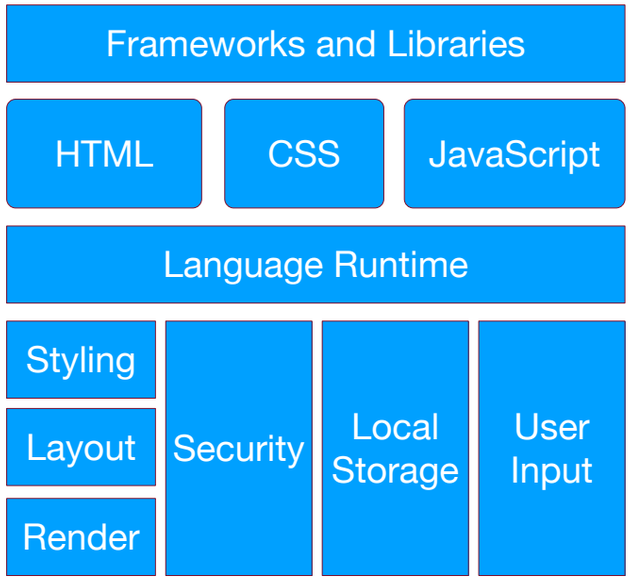


Application

- Parallelize browser computation

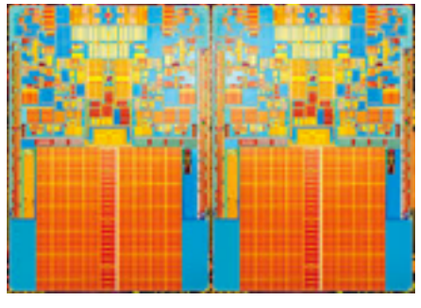
Architecture

Traditional Approach



Application

- Parallelize browser computation

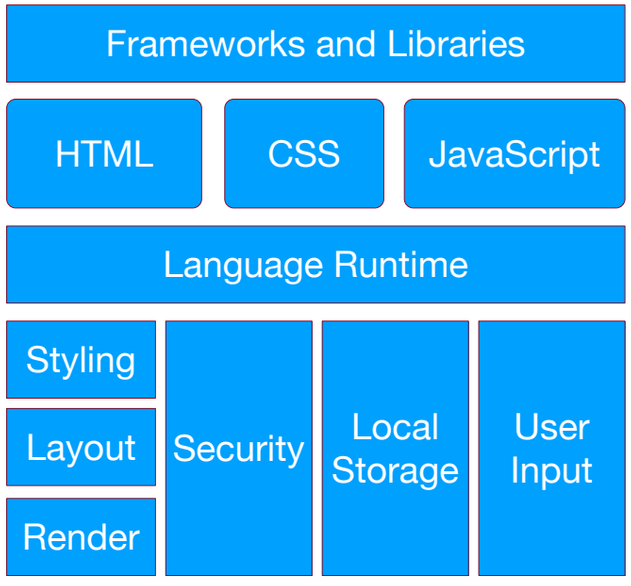


Architecture

- Voltage/frequency scaling on general-purpose processors

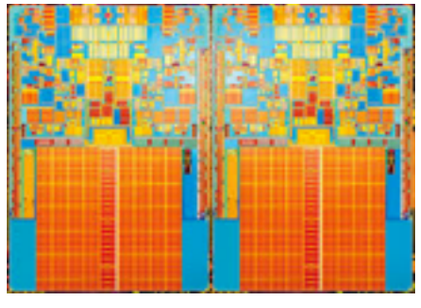
Traditional Approach

Inputs



Application

- Parallelize browser computation



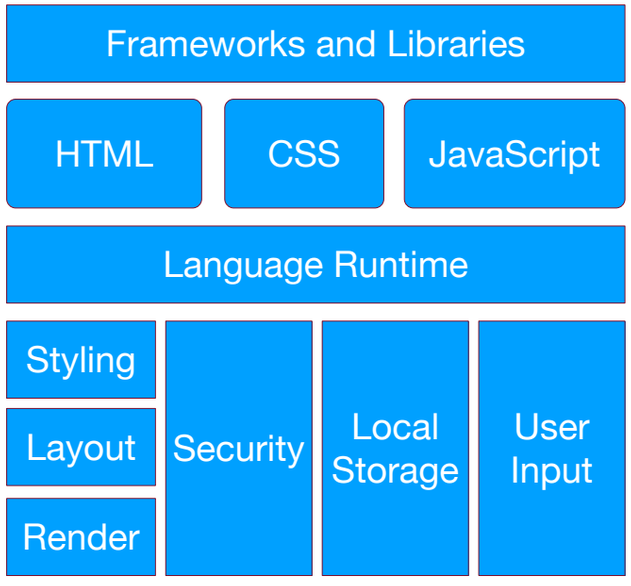
Architecture

- Voltage/frequency scaling on general-purpose processors

Traditional Approach

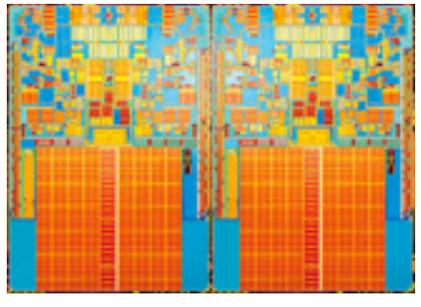
Inputs

- Ignored!



Application

- Parallelize browser computation



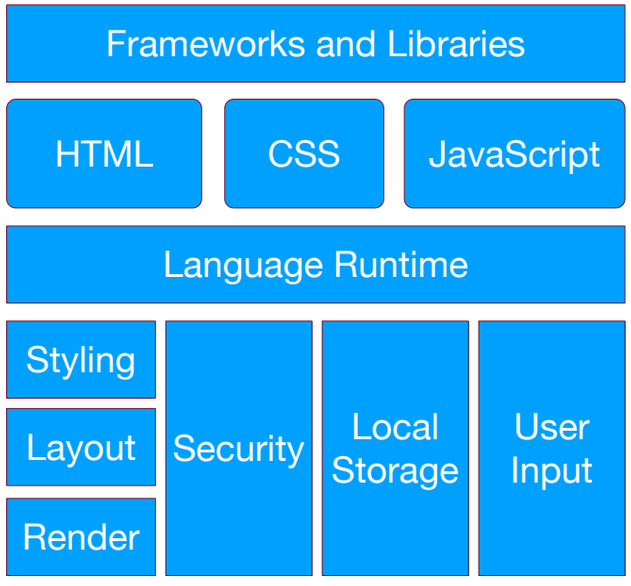
Architecture

- Voltage/frequency scaling on general-purpose processors

Traditional Approach

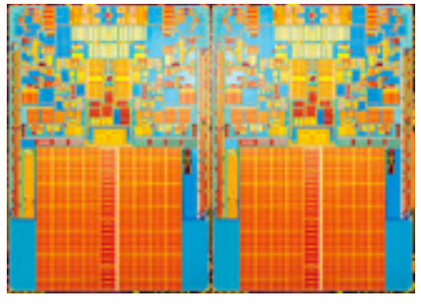
Inputs

- Ignored!



Application

- Parallelize browser computation



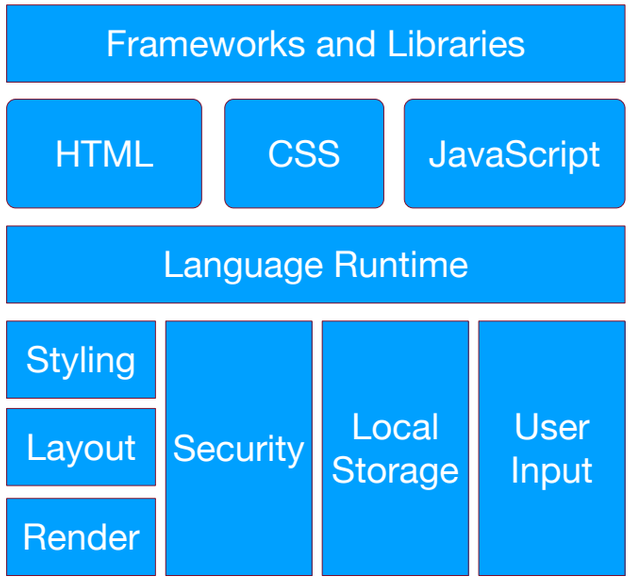
Architecture

- End of Dennard Scaling!
- Diminishing return

My Approach

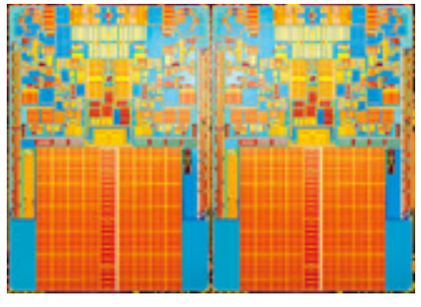
Inputs

- Ignored!



Application

- Parallelize browser computation

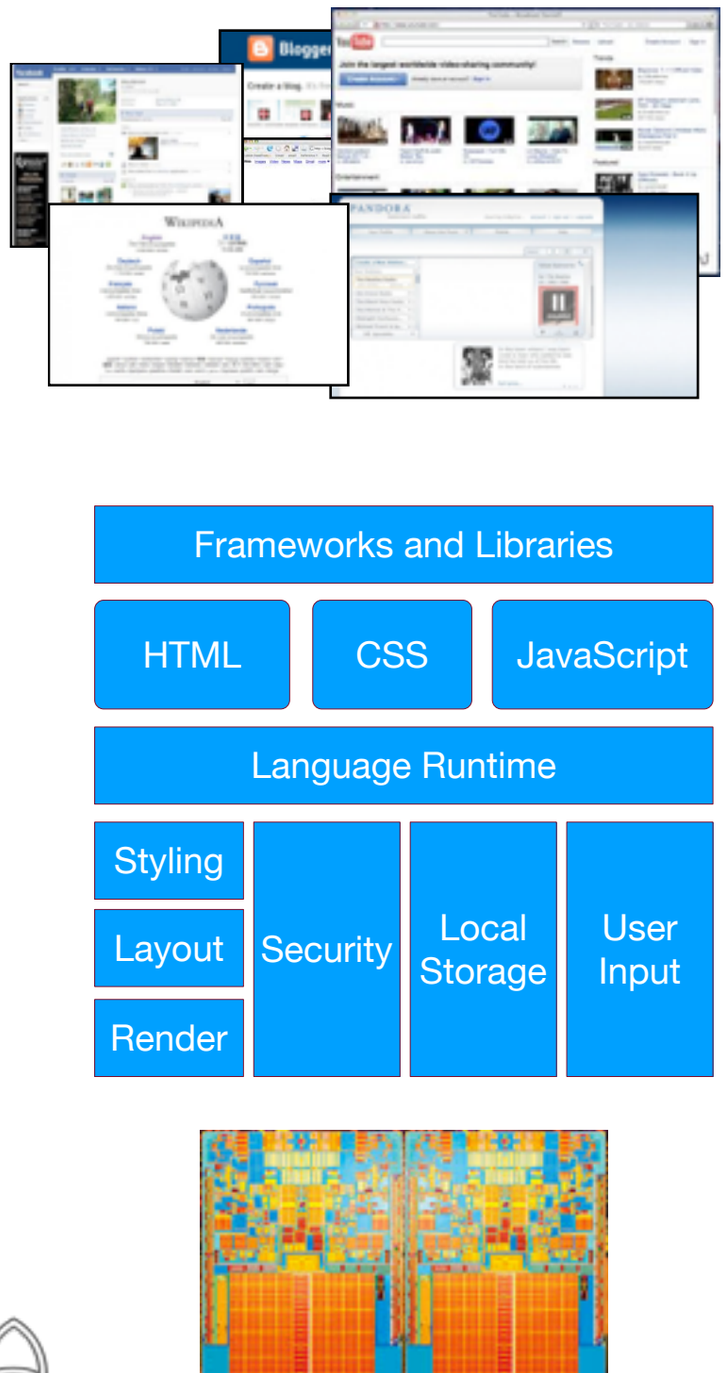


Architecture

WebCore

Web-specific Architecture

My Approach



Inputs

- Lost page-level diversity
- Lost user QoS requirements

Application

- Parallelize browser computation

Architecture

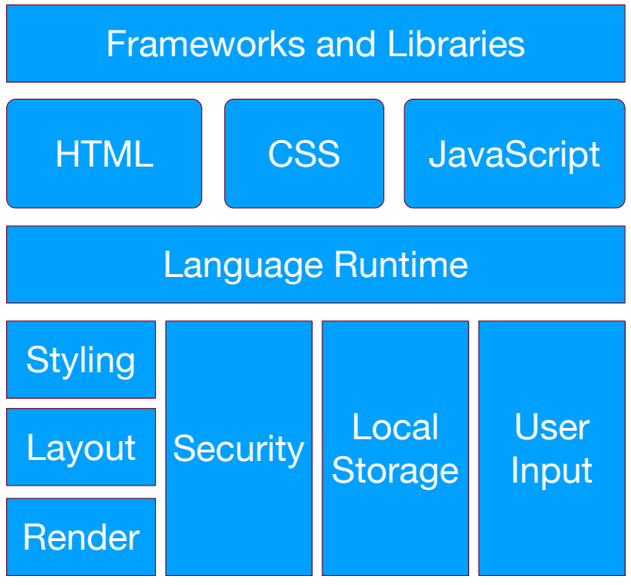
WebCore

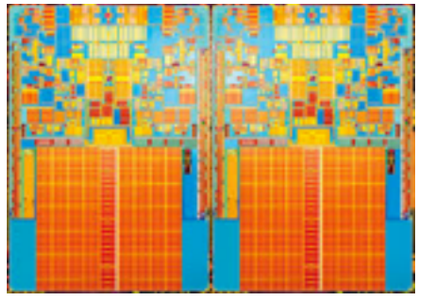
Web-specific Architecture

My Approach

Application

- Lost page-level diversity
- Lost user QoS requirements





Architecture

- Parallelize browser computation

WebCore

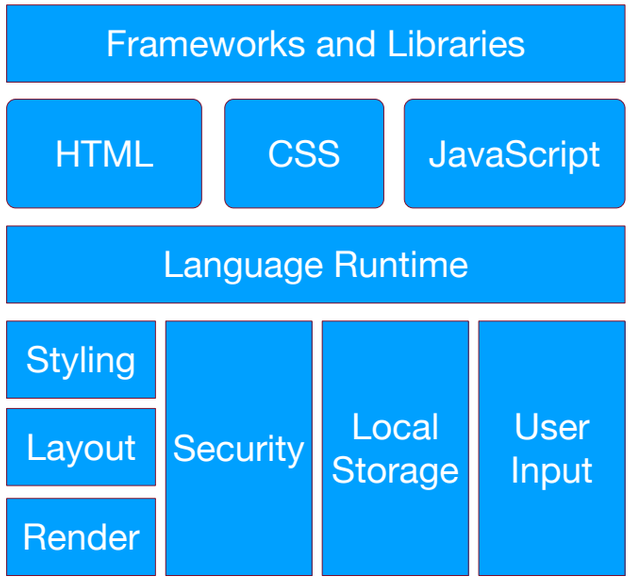
Web-specific Architecture

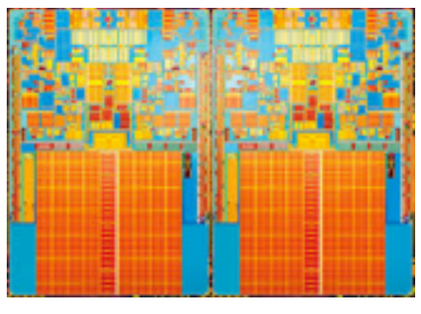
My Approach

Application

GreenWeb

Language Extensions





Architecture

WebCore

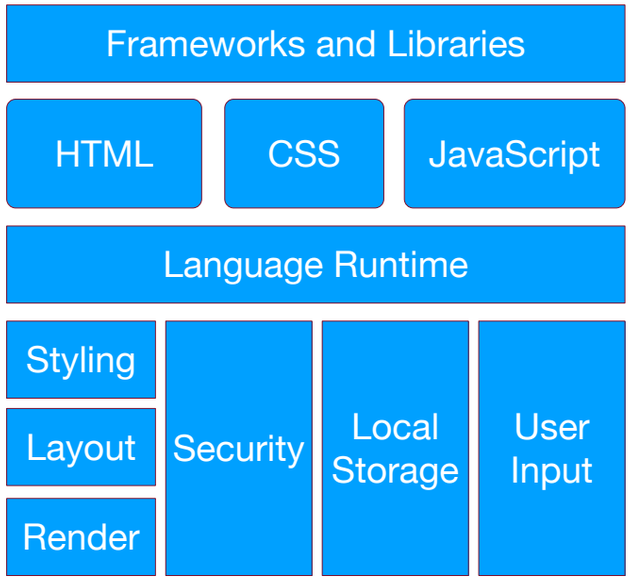
Web-specific Architecture

My Approach

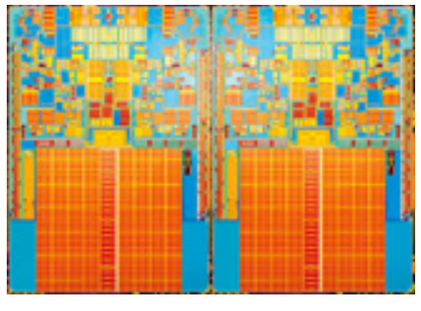
Application

GreenWeb

Language Extensions



Runtime



Architecture

WebCore

Web-specific Architecture

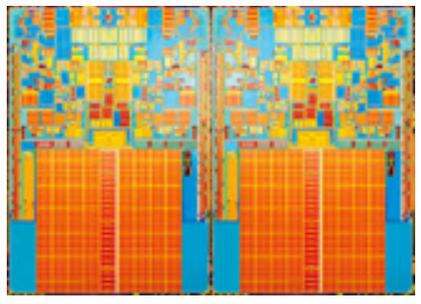
My Approach

Frameworks and Libraries

HTML | CSS | JavaScript

Language Runtime

Styling | Layout | Render | Security | Local Storage | User Input



Application

GreenWeb

Language Extensions

Runtime

Architecture

WebCore

Web-specific Architecture

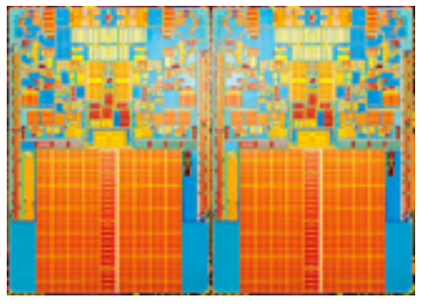
My Approach

Frameworks and Libraries

HTML | CSS | JavaScript

Language Runtime

Styling | Layout | Render | Security | Local Storage | User Input



Application

GreenWeb

Language Extensions

Runtime

Architecture

WebCore

Web-specific Architecture

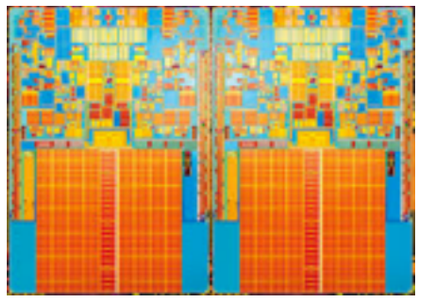
My Approach

Frameworks and Libraries

HTML | CSS | JavaScript

Language Runtime

Styling | Layout | Render | Security | Local Storage | User Input



Application

Runtime

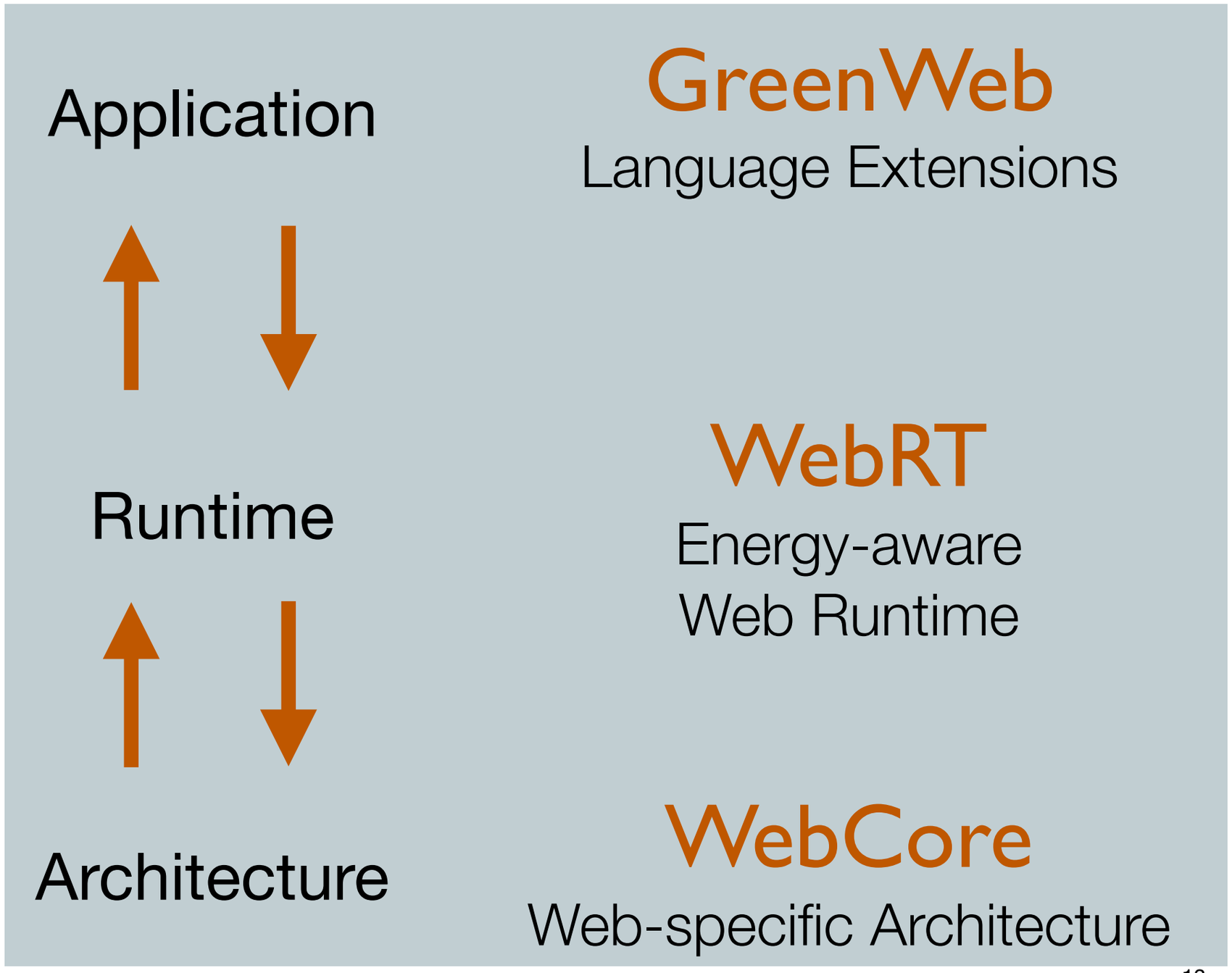
Architecture

GreenWeb
Language Extensions

WebRT
Energy-aware
Web Runtime

WebCore
Web-specific Architecture

My Approach


Application
GreenWeb
Language Extensions
Runtime
WebRT
Energy-aware
Web Runtime
Architecture
WebCore
Web-specific Architecture

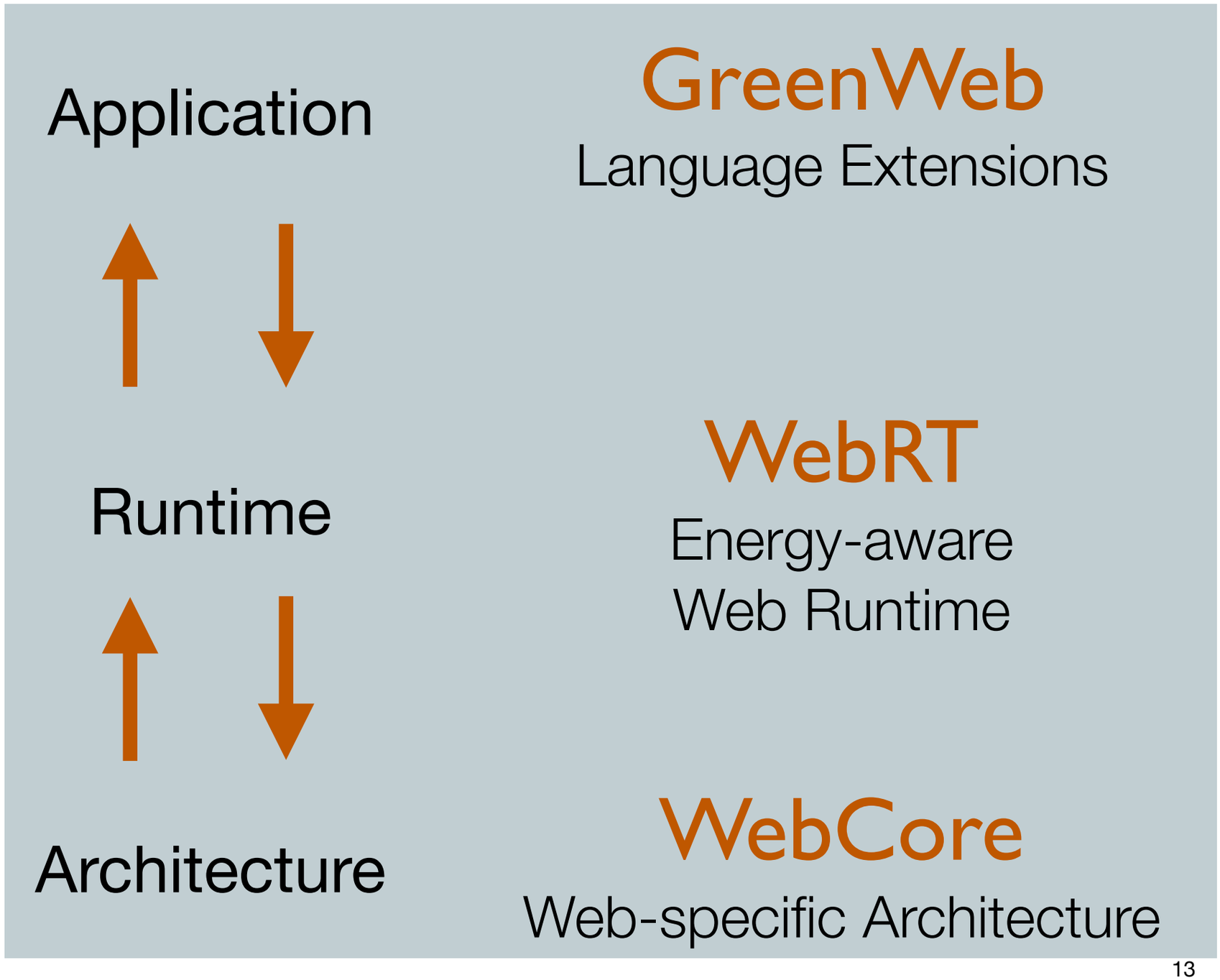
My Approach

My Dissertation Work

[PLDI 2016]

[HPCA 2013]
[HPCA 2015]
[CAL 2014]
(Best of CAL)

[ISCA 2014]



Thesis Statement

Thesis Statement

Future mobile Web systems can achieve energy-efficiency without sacrificing responsiveness by incorporating:

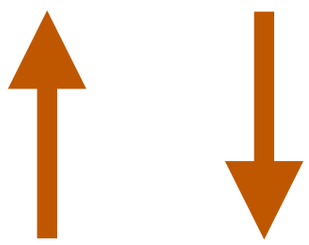
Thesis Statement

Future mobile Web systems can achieve energy-efficiency without sacrificing responsiveness by incorporating:

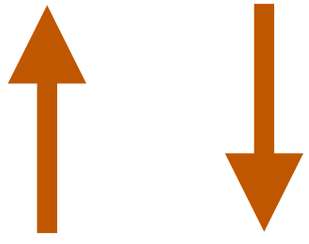
- Programming language annotations to convey user QoS information
- Runtime scheduling mechanisms to exploit heterogeneous hardware
- Hardware accelerators specialized for the key computation kernel

My Approach

Application



Runtime



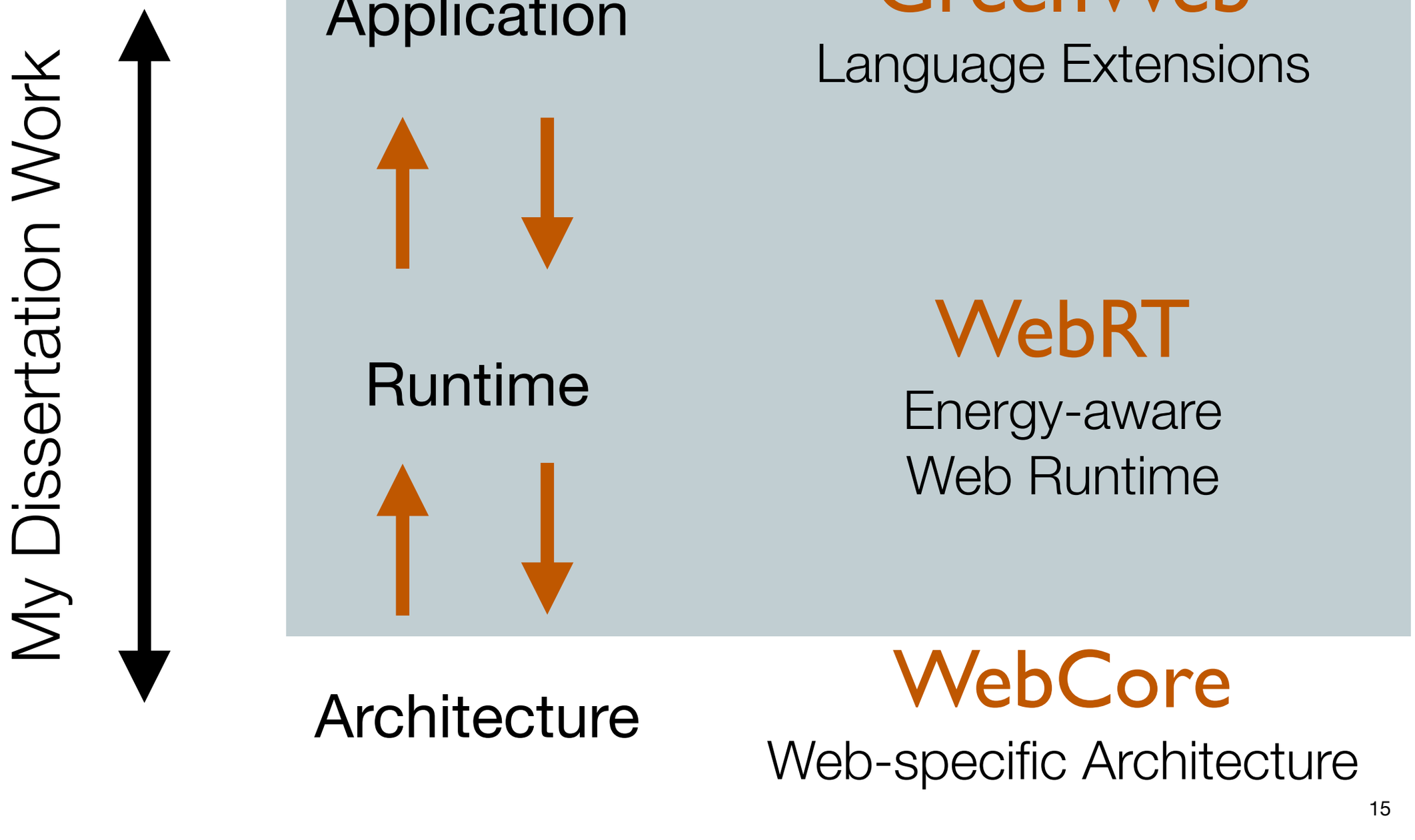
Architecture

GreenWeb
Language Extensions

WebRT
Energy-aware
Web Runtime

WebCore
Web-specific Architecture

My Approach

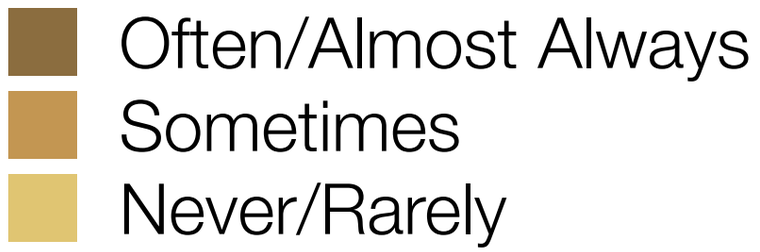

My Dissertation Work
Application
Runtime
Architecture
GreenWeb
Language Extensions
WebRT
Energy-aware
Web Runtime
WebCore
Web-specific Architecture

Energy Concern Among Mobile Developers

[ICSE 2016] Manotas et al., “An Empirical Study of Practitioners’ Perspectives on Green Software Engineering”

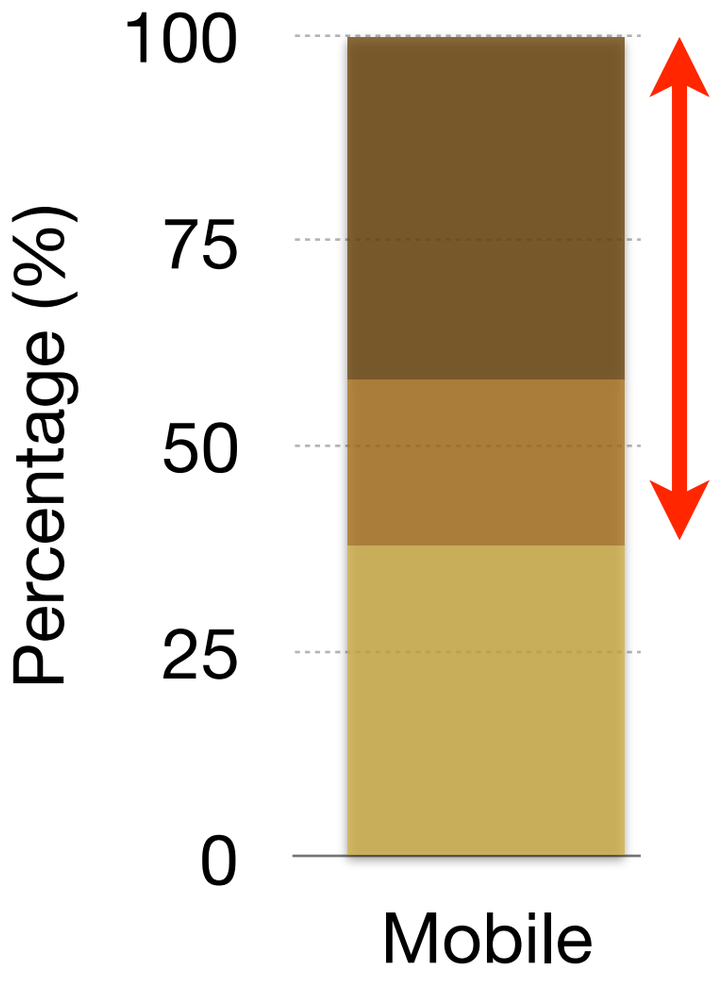
Energy Concern Among Mobile Developers

"My applications have requirements about energy usage."



[ICSE 2016] Manotas et al., "An Empirical Study of Practitioners' Perspectives on Green Software Engineering"

Energy Concern Among Mobile Developers

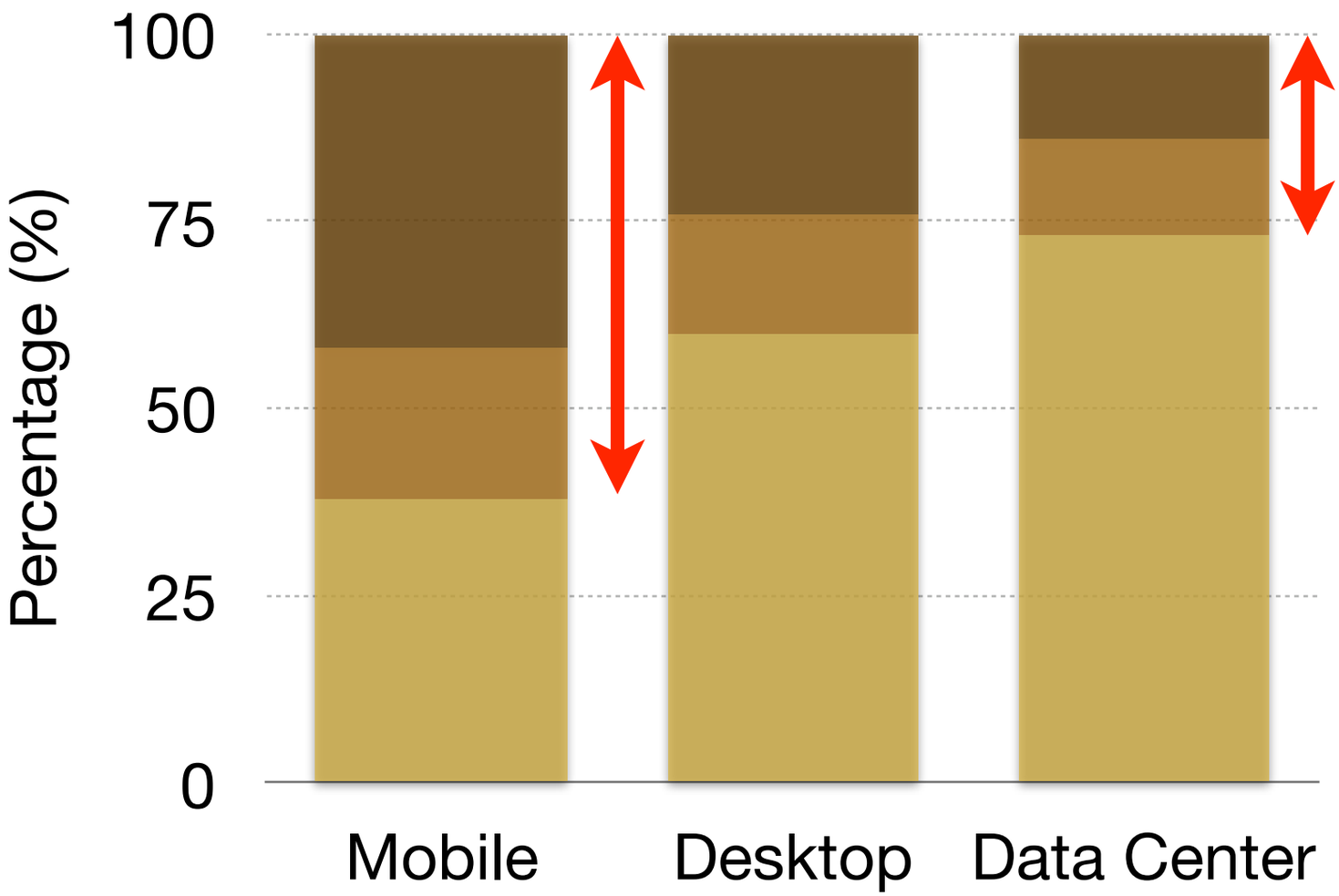


"My applications have requirements about energy usage."

- Often/Almost Always
- Sometimes
- Never/Rarely

[ICSE 2016] Manotas et al., "An Empirical Study of Practitioners' Perspectives on Green Software Engineering"

Energy Concern Among Mobile Developers



"My applications have requirements about energy usage."

- Often/Almost Always
- Sometimes
- Never/Rarely

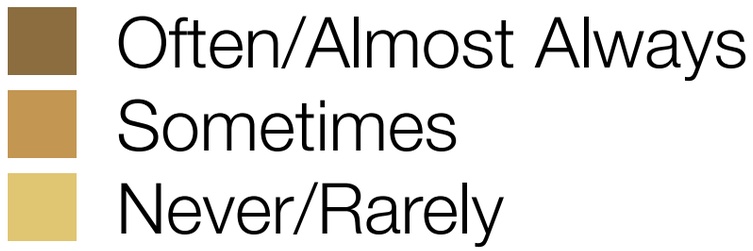
[ICSE 2016] Manotas et al., "An Empirical Study of Practitioners' Perspectives on Green Software Engineering"

Developers are Willing to Make Trade-offs

[ICSE 2016] Manotas et al., “An Empirical Study of Practitioners’ Perspectives on Green Software Engineering”

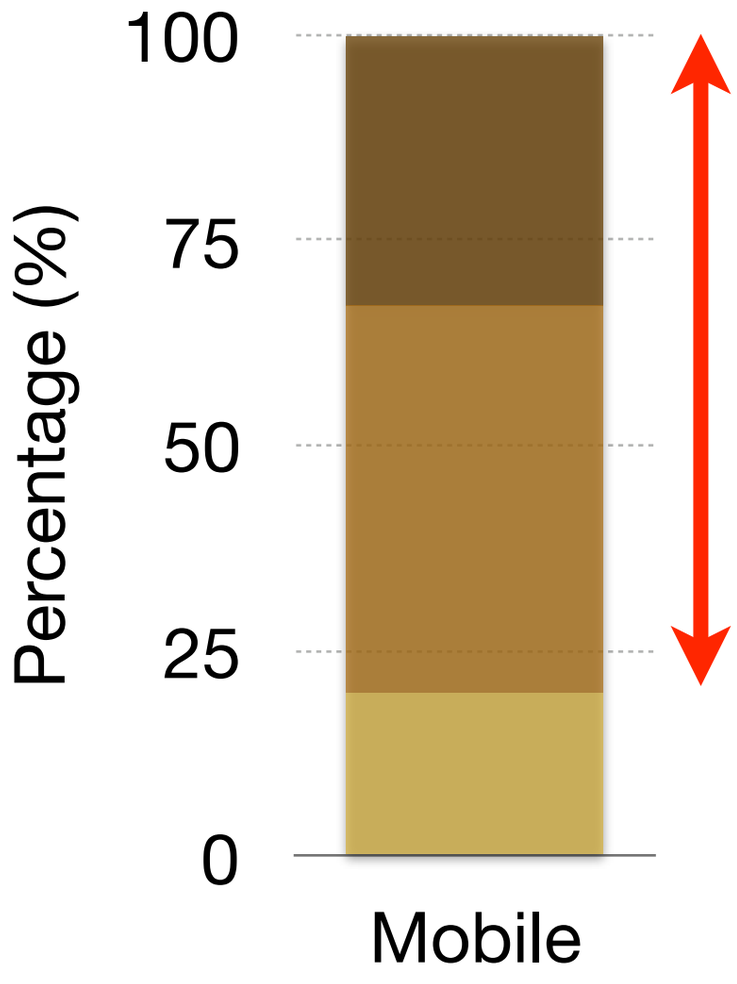
Developers are Willing to Make Trade-offs

"I'm willing to sacrifice performance, etc. for reduced energy usage."



[ICSE 2016] Manotas et al., "An Empirical Study of Practitioners' Perspectives on Green Software Engineering"

Developers are Willing to Make Trade-offs



"I'm willing to sacrifice performance, etc. for reduced energy usage."

Often/Almost Always
Sometimes
Never/Rarely

[ICSE 2016] Manotas et al., "An Empirical Study of Practitioners' Perspectives on Green Software Engineering"

Energy-efficiency

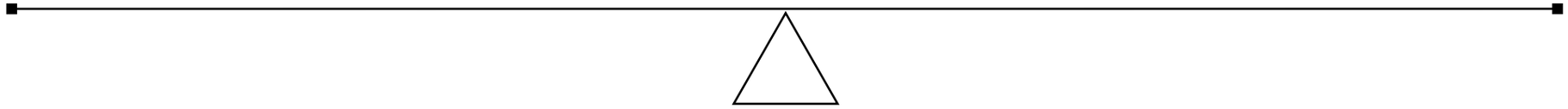
Quality-of-service

Energy-efficiency

Quality-of-service

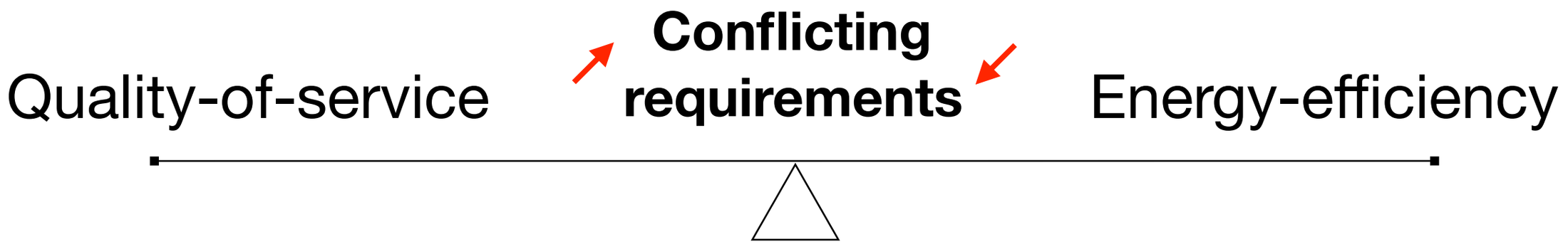
Conflicting requirements

Energy-efficiency



GreenWeb

Programming language support for balancing energy-efficiency and QoS in mobile Web computing



GreenWeb

Programming language support for balancing energy-efficiency and QoS in mobile Web computing

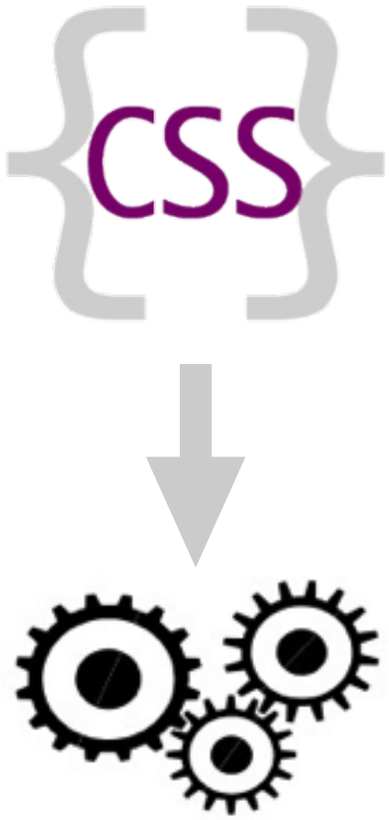
GreenWeb

Programming language support for balancing energy-efficiency and QoS in mobile Web computing

GreenWeb: Language for Energy-Efficiency

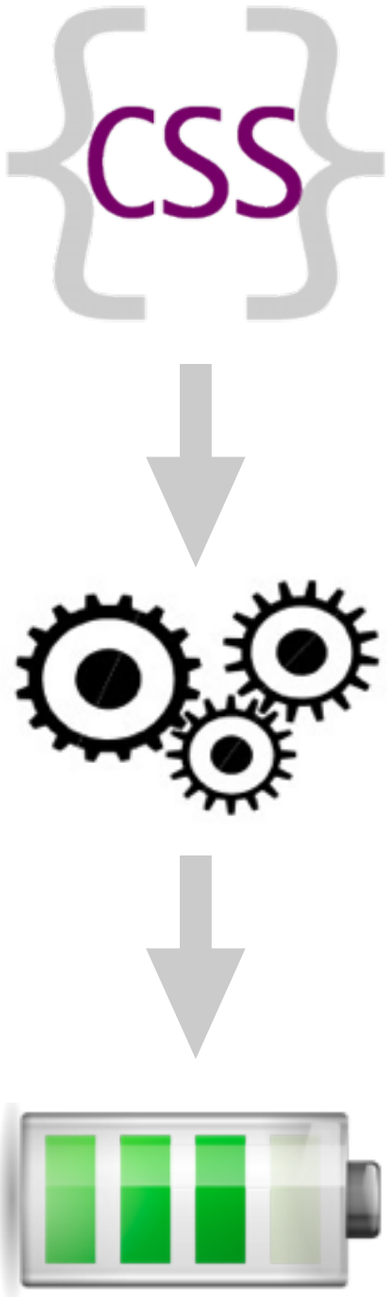
- **Language abstractions** for expressing QoS

GreenWeb: Language for Energy-Efficiency



- **Language abstractions** for expressing QoS
- **Runtime** that saves energy while meeting the QoS constraints

GreenWeb: Language for Energy-Efficiency



- **Language abstractions** for expressing QoS

- **Runtime** that saves energy while meeting the QoS constraints

- **Result** in 60% energy savings on real hardware/software implementations

GreenWeb: Language for Energy-Efficiency

- **Language abstractions** for expressing QoS

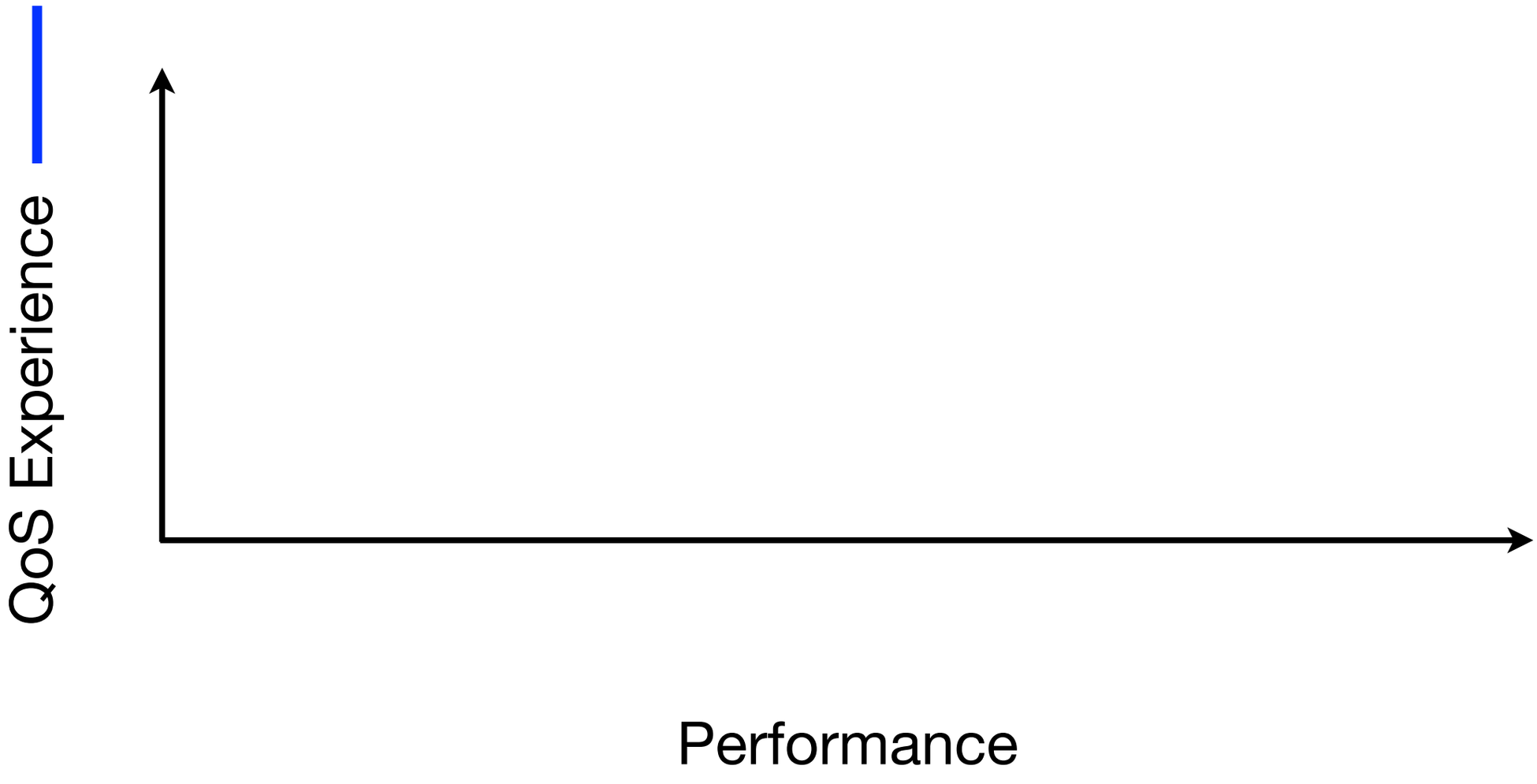
- Runtime
  the QoS constraints

- Result
  hardware/software implementations

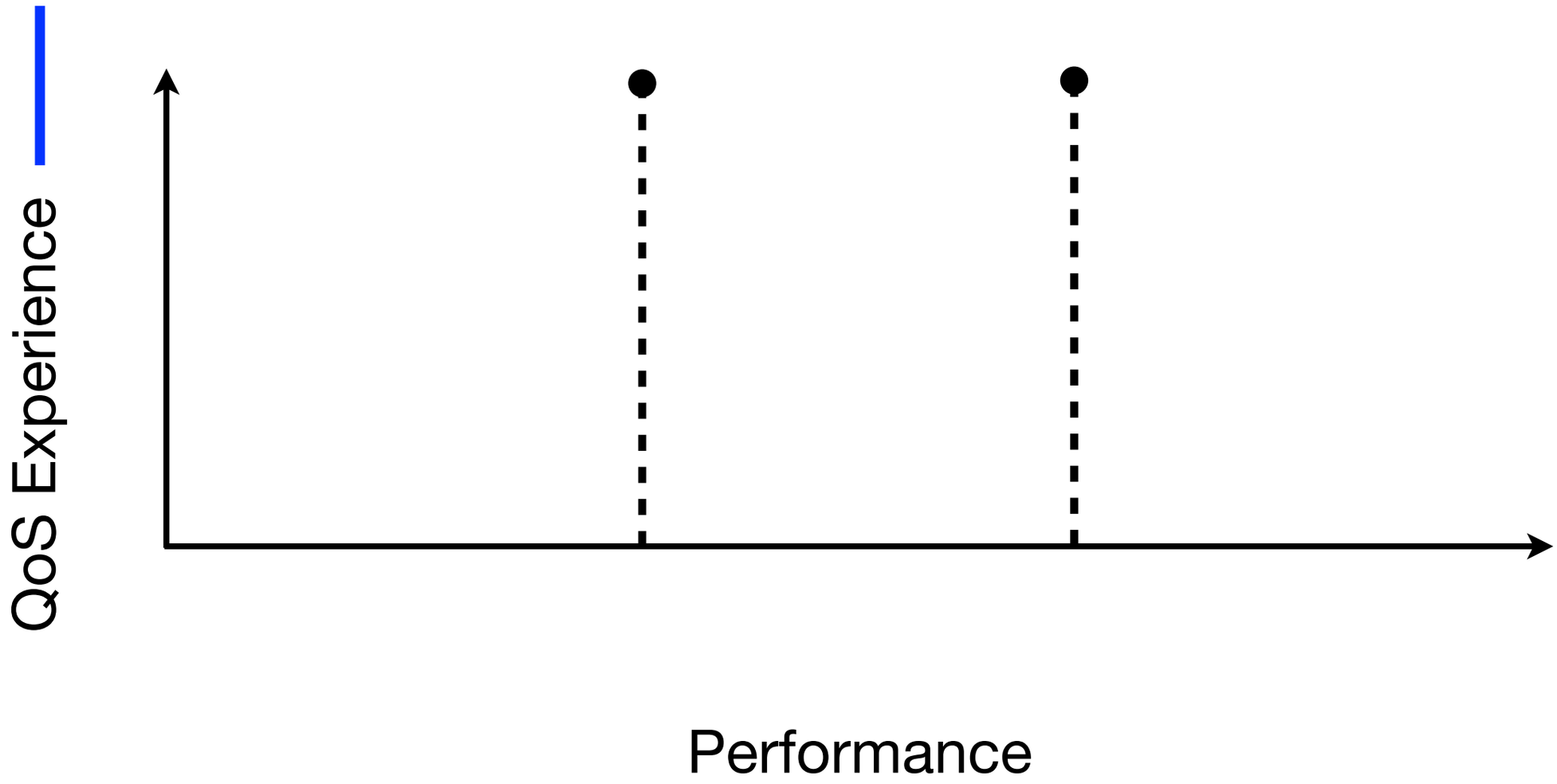
What is QoS in mobile Web?

Understanding Mobile Web QoS

Understanding Mobile Web QoS

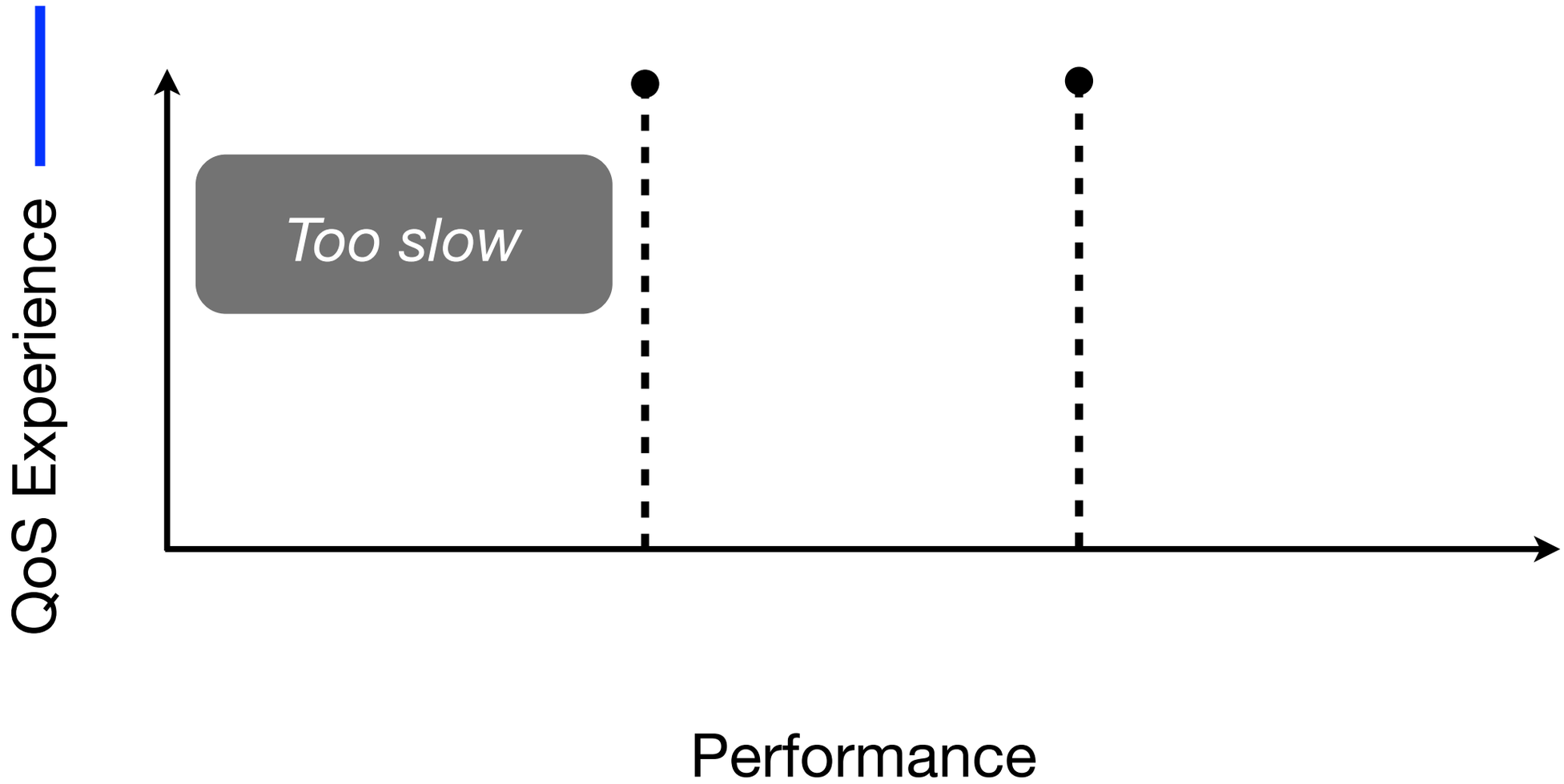


Understanding Mobile Web QoS



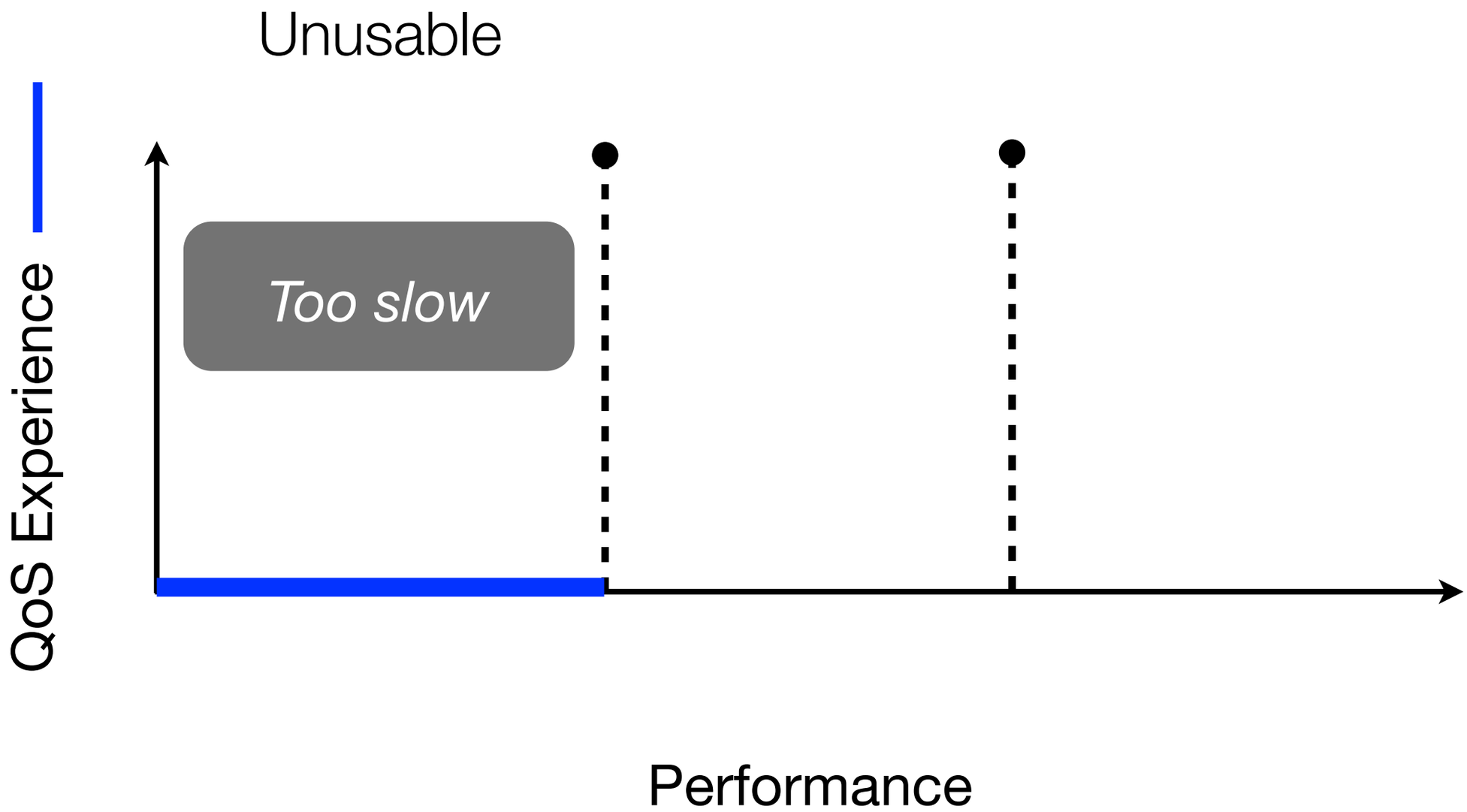
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



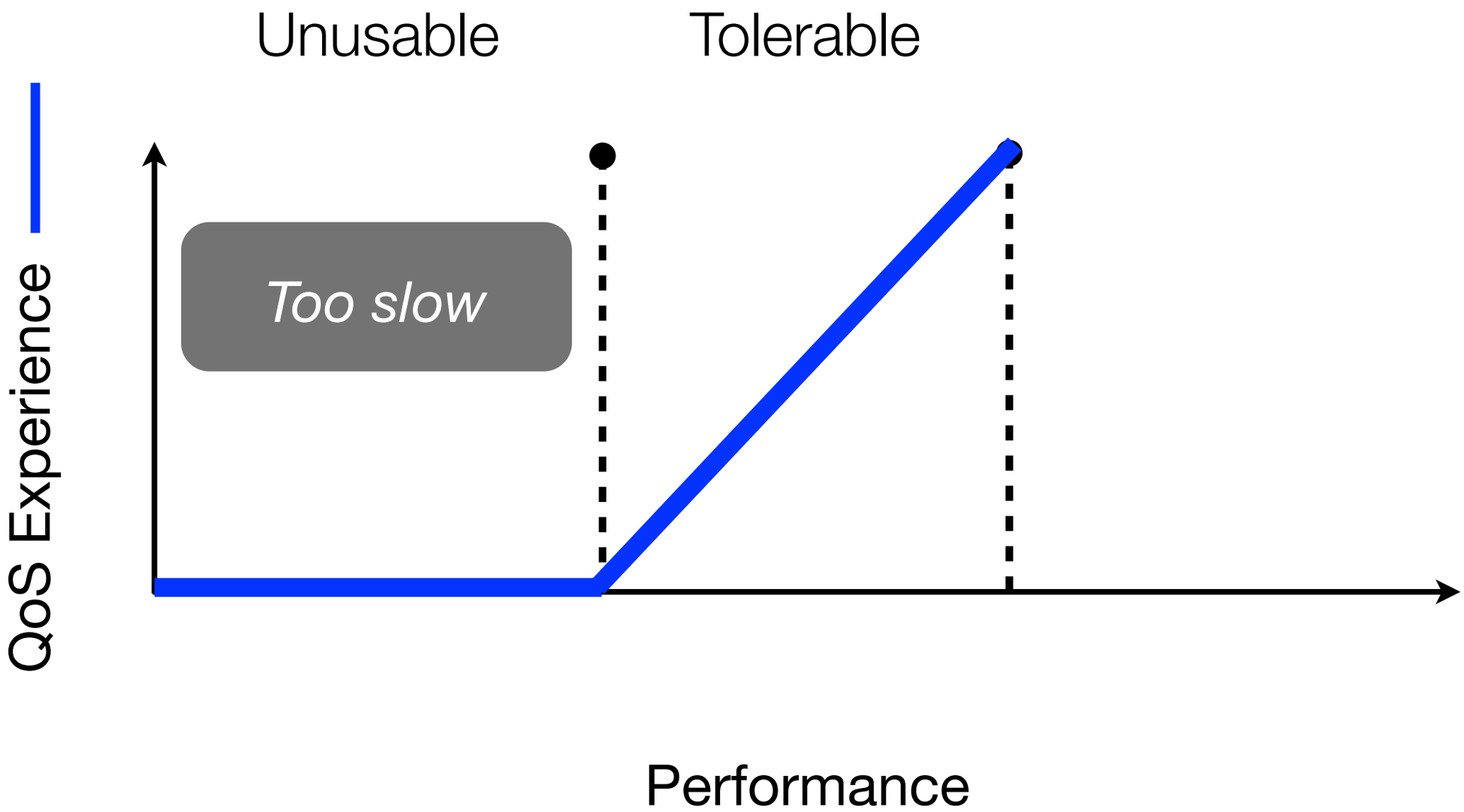
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



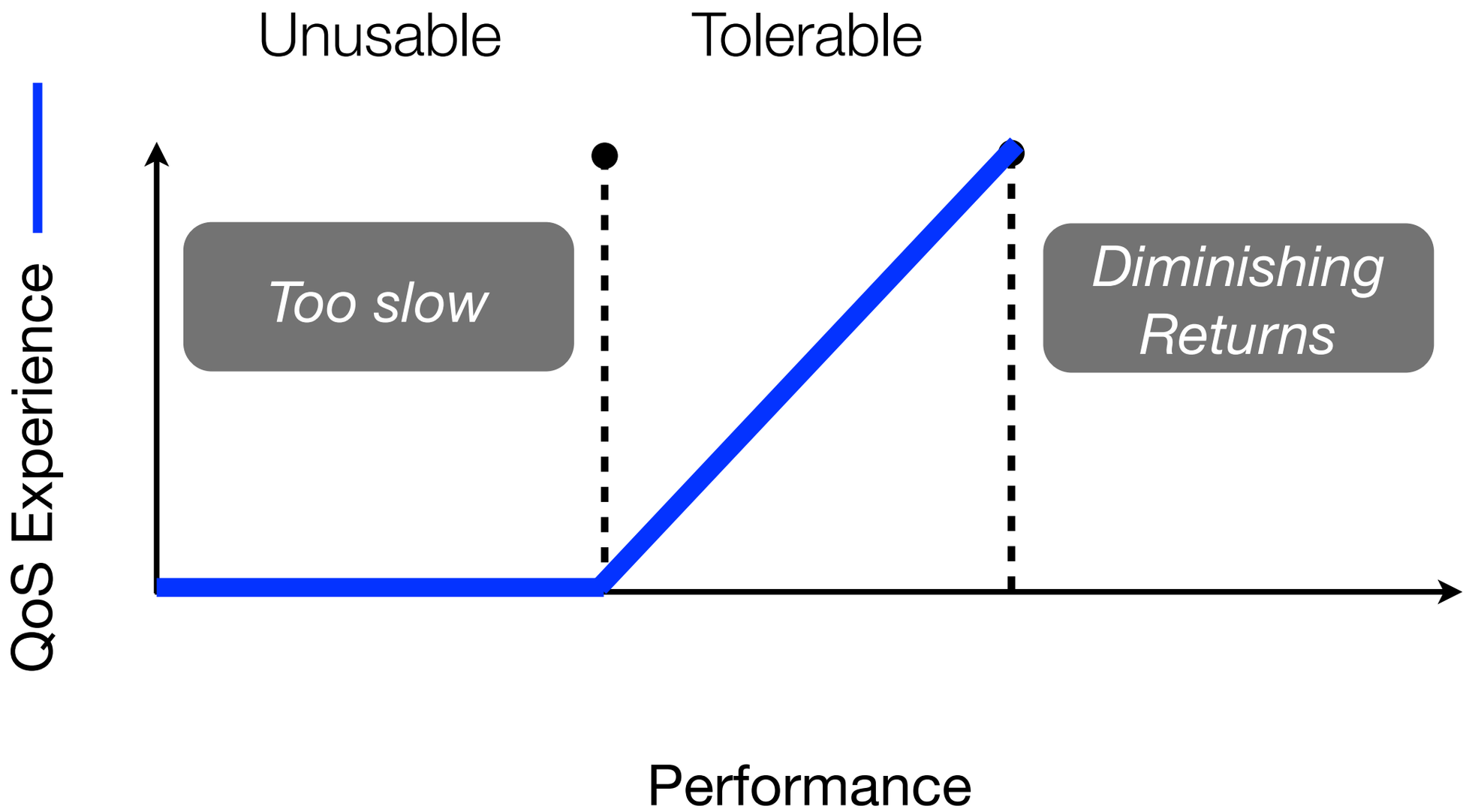
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



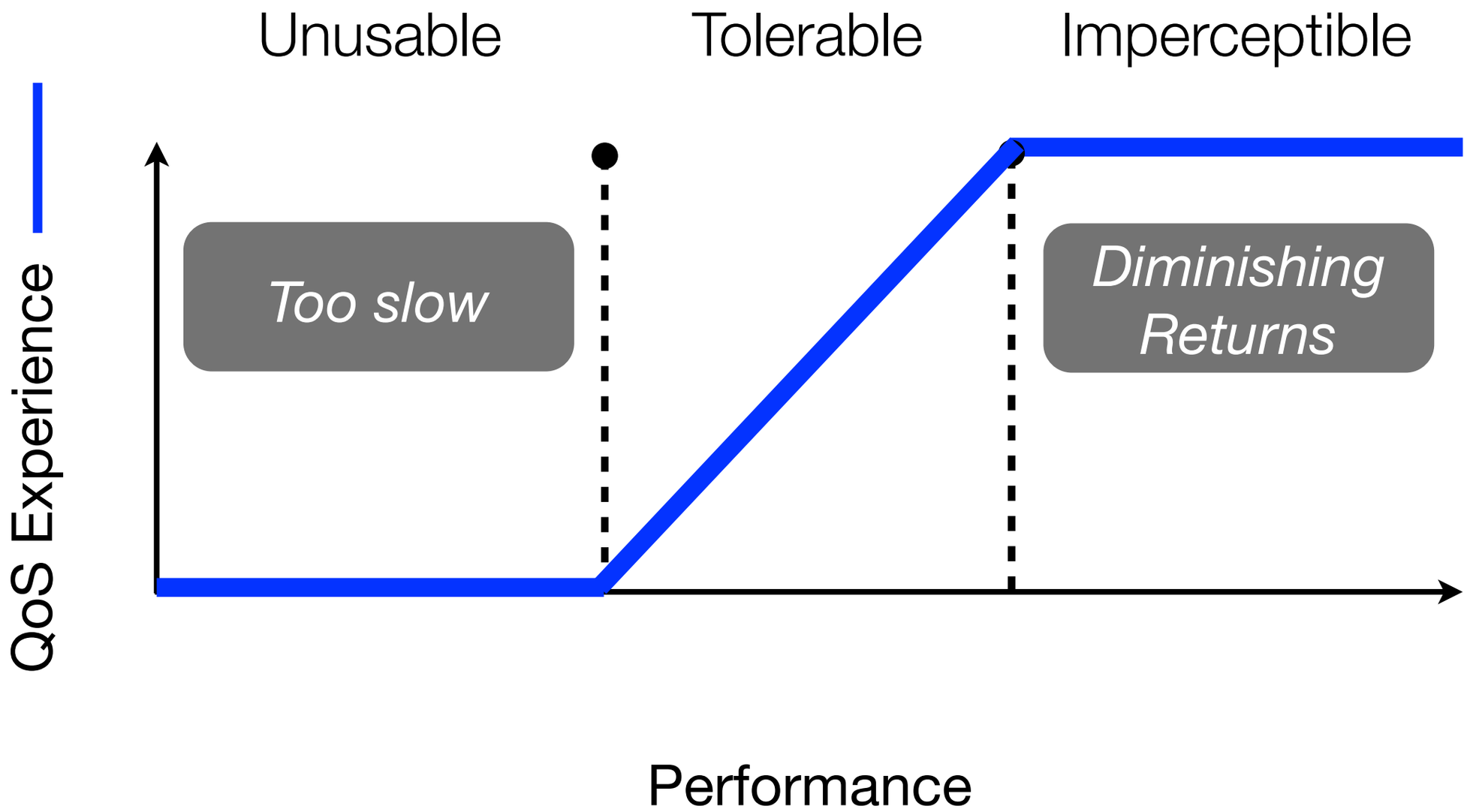
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



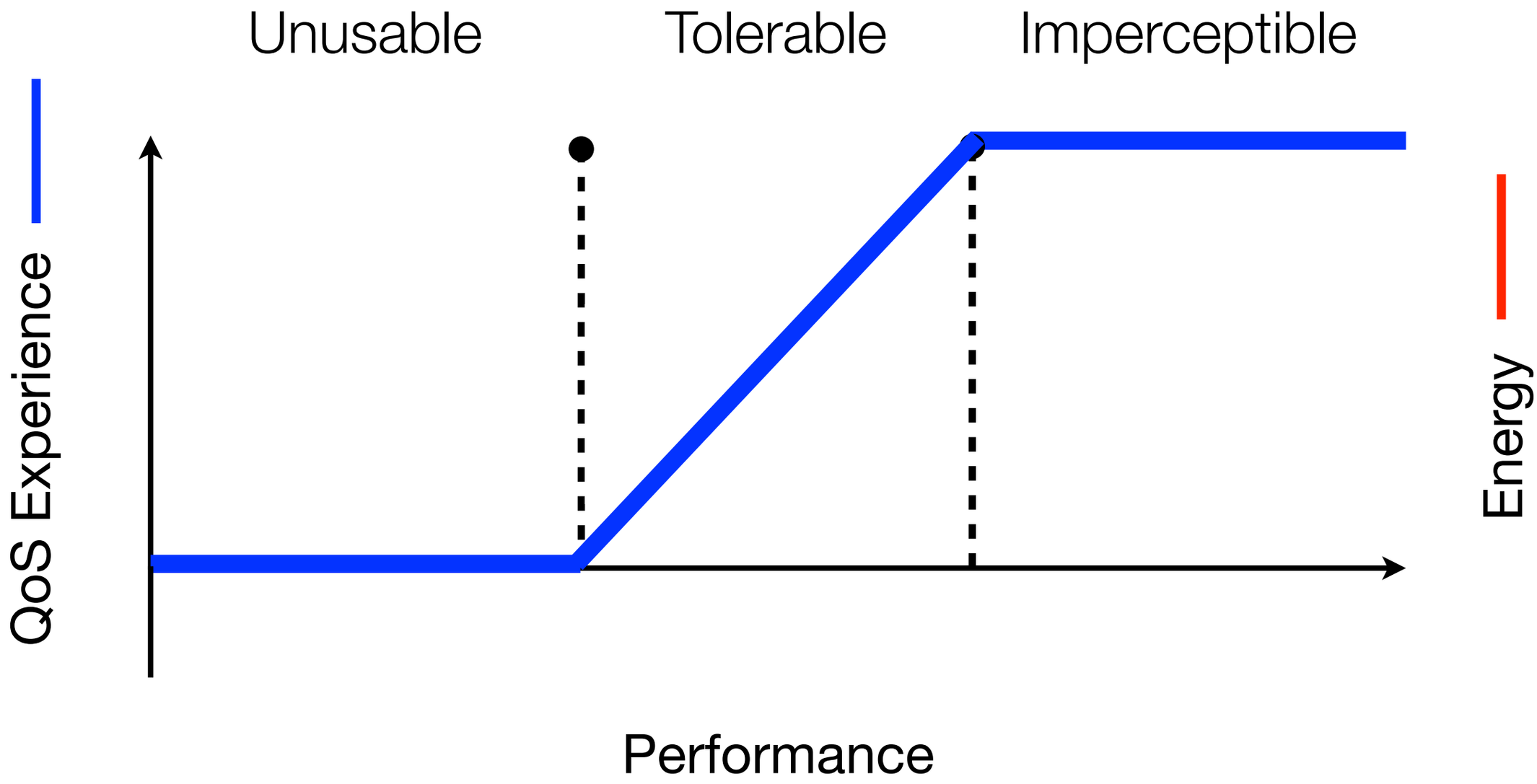
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



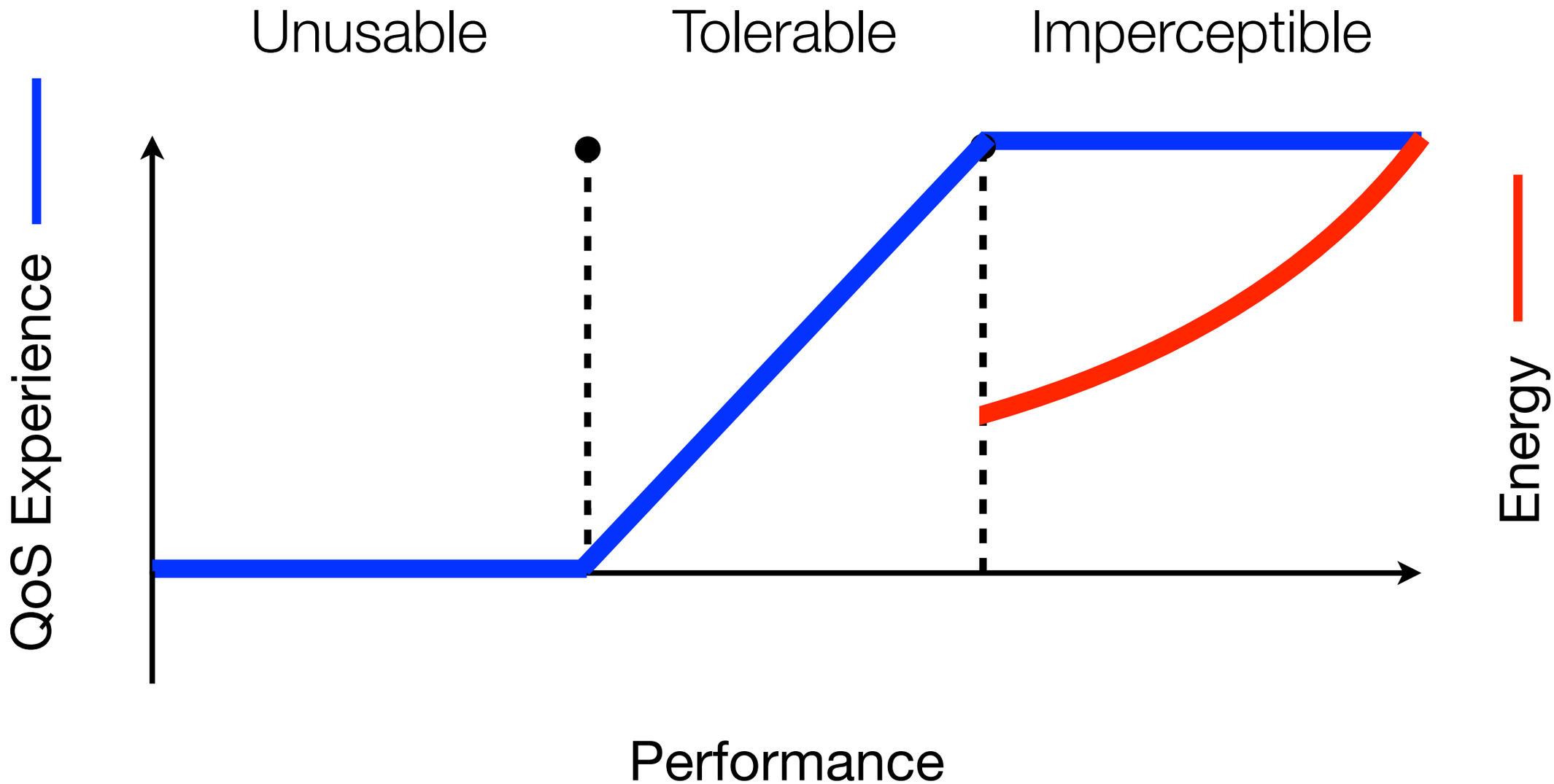
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



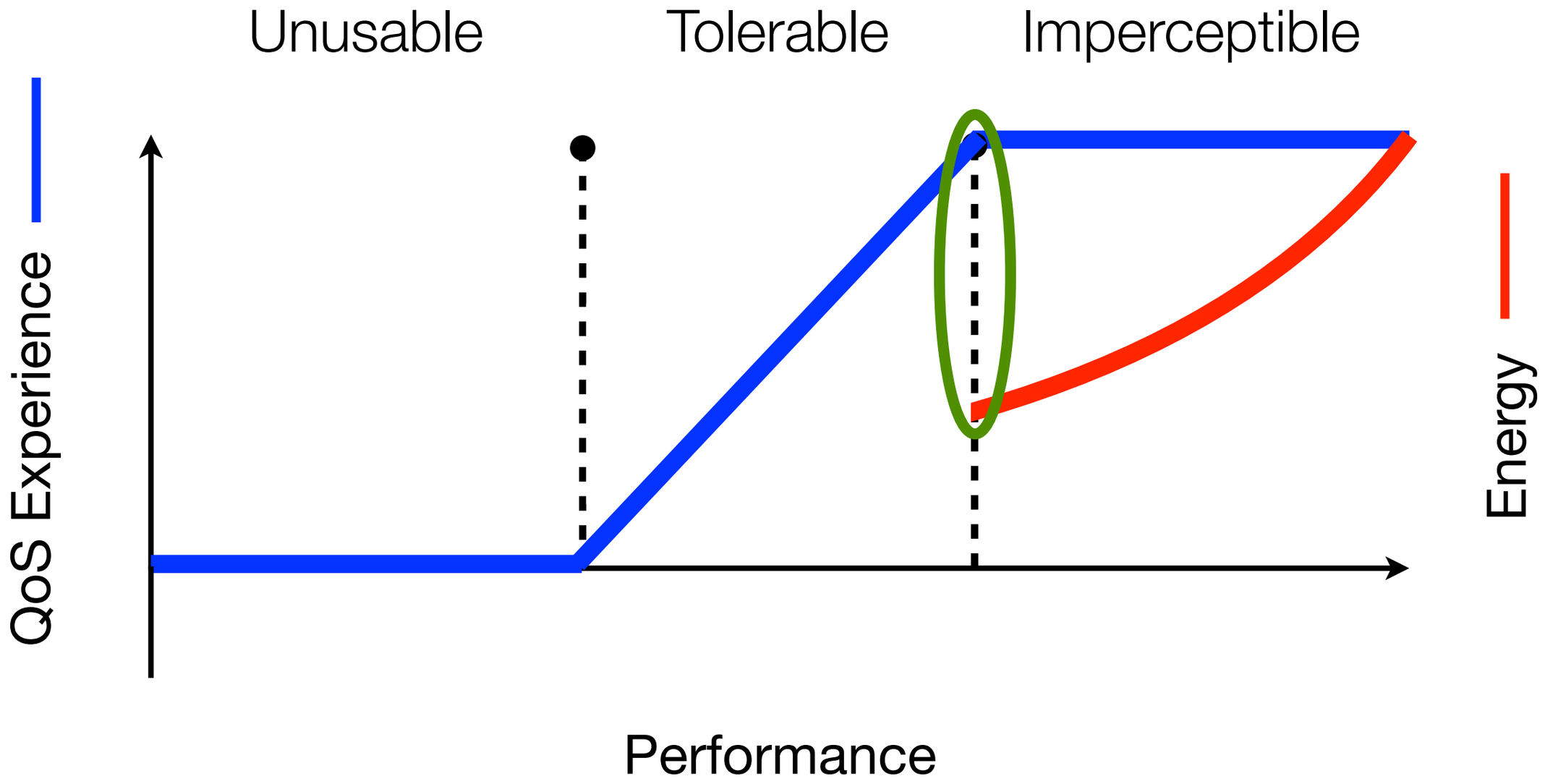
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



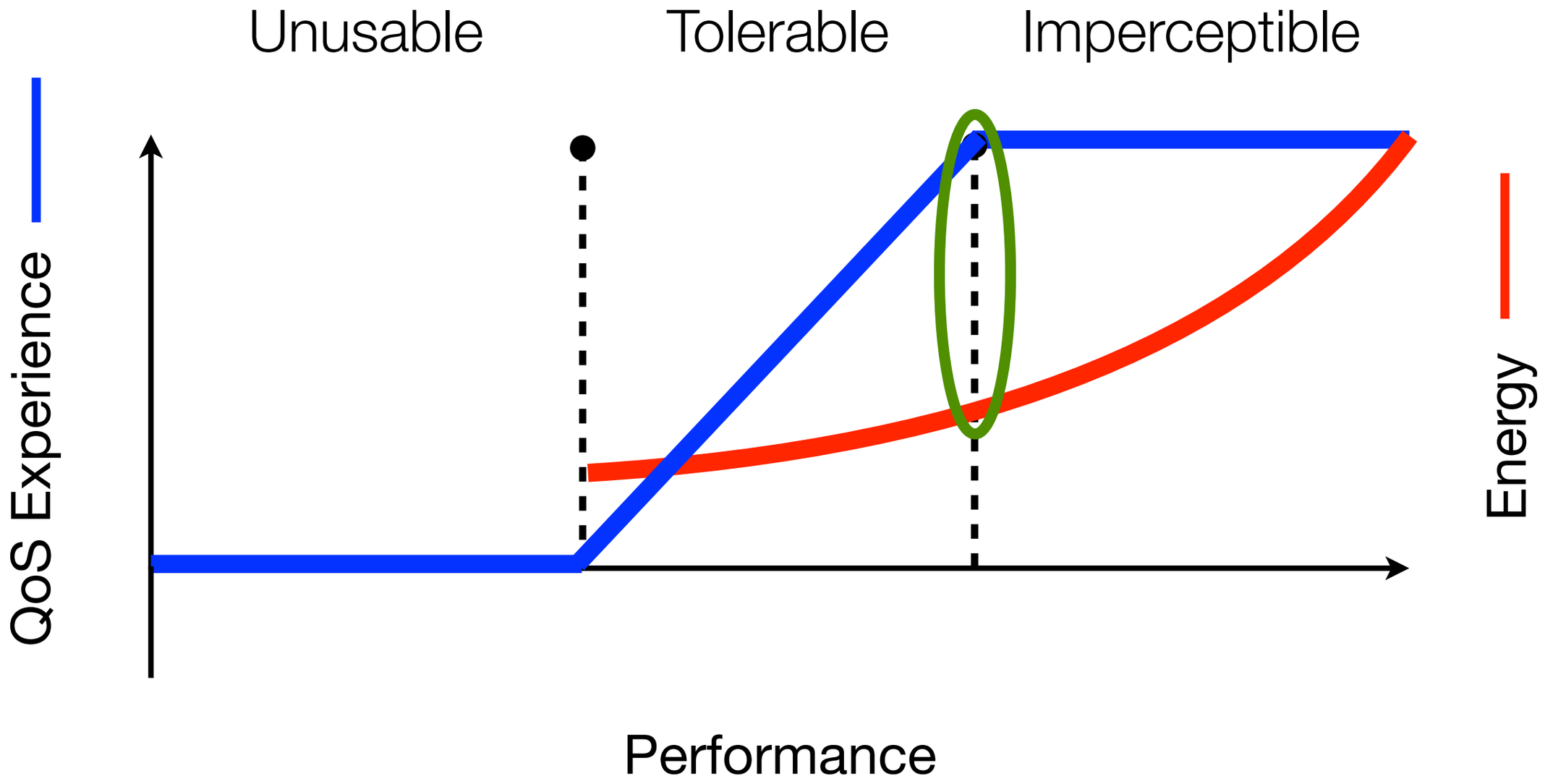
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



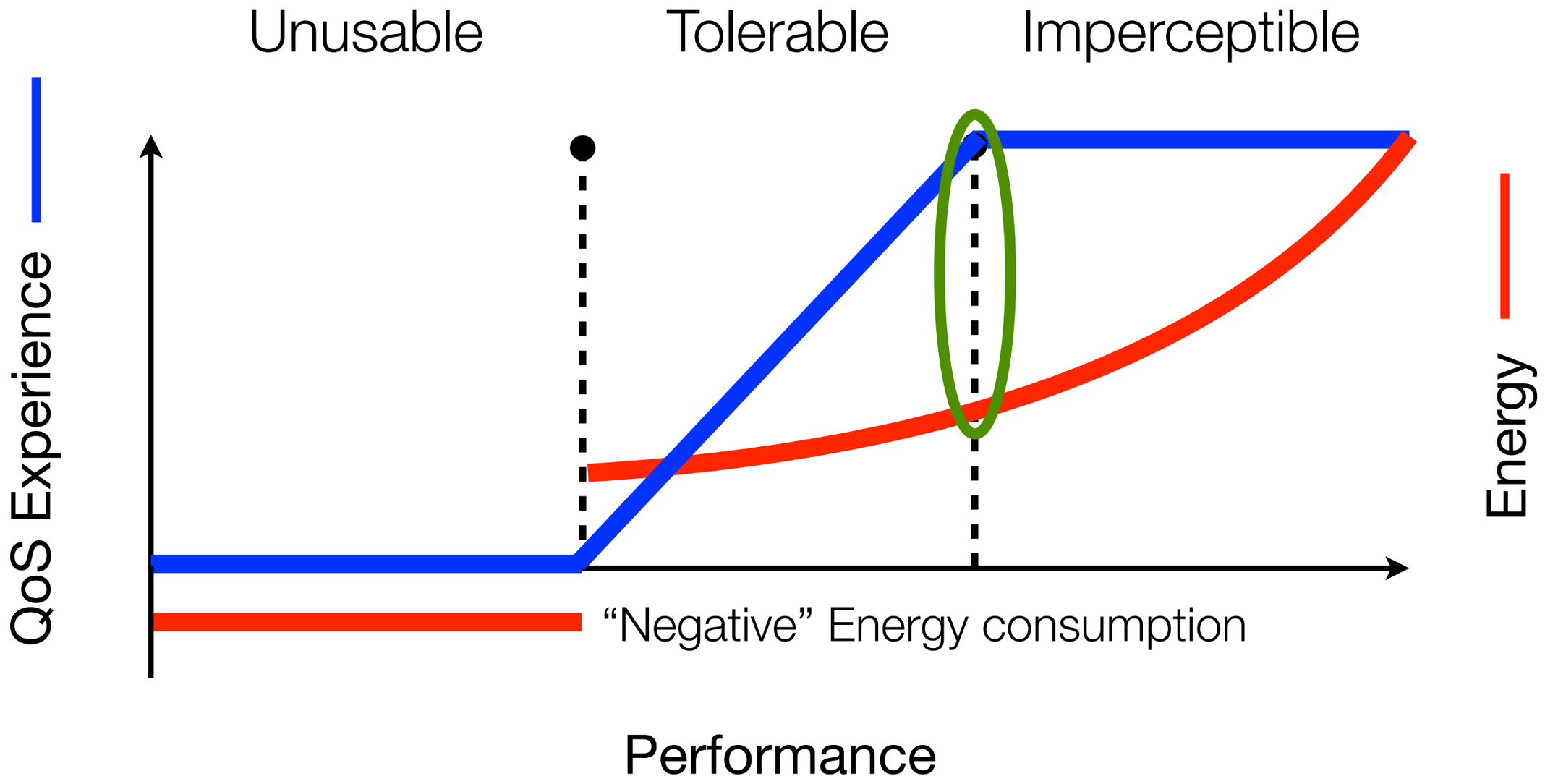
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



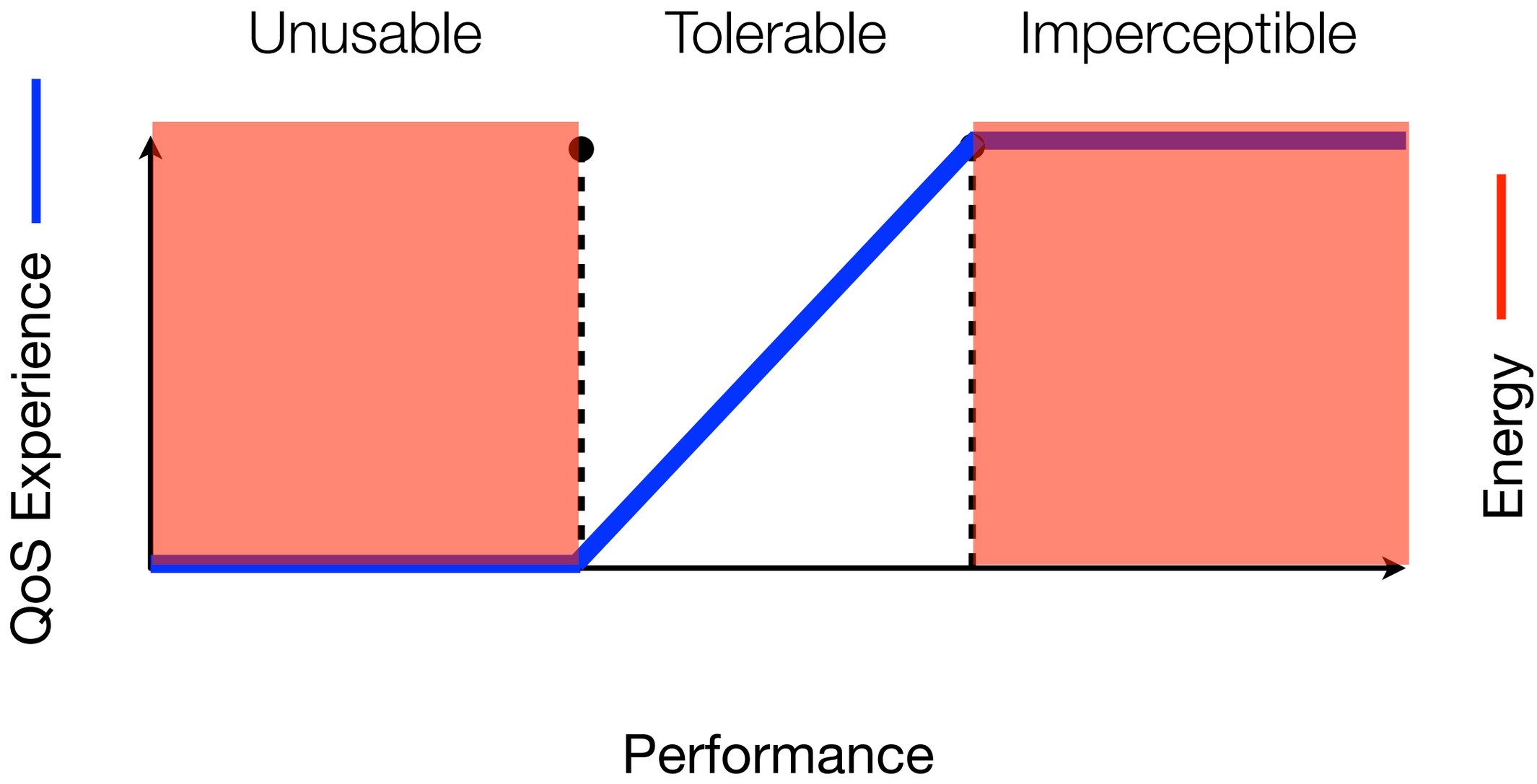
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



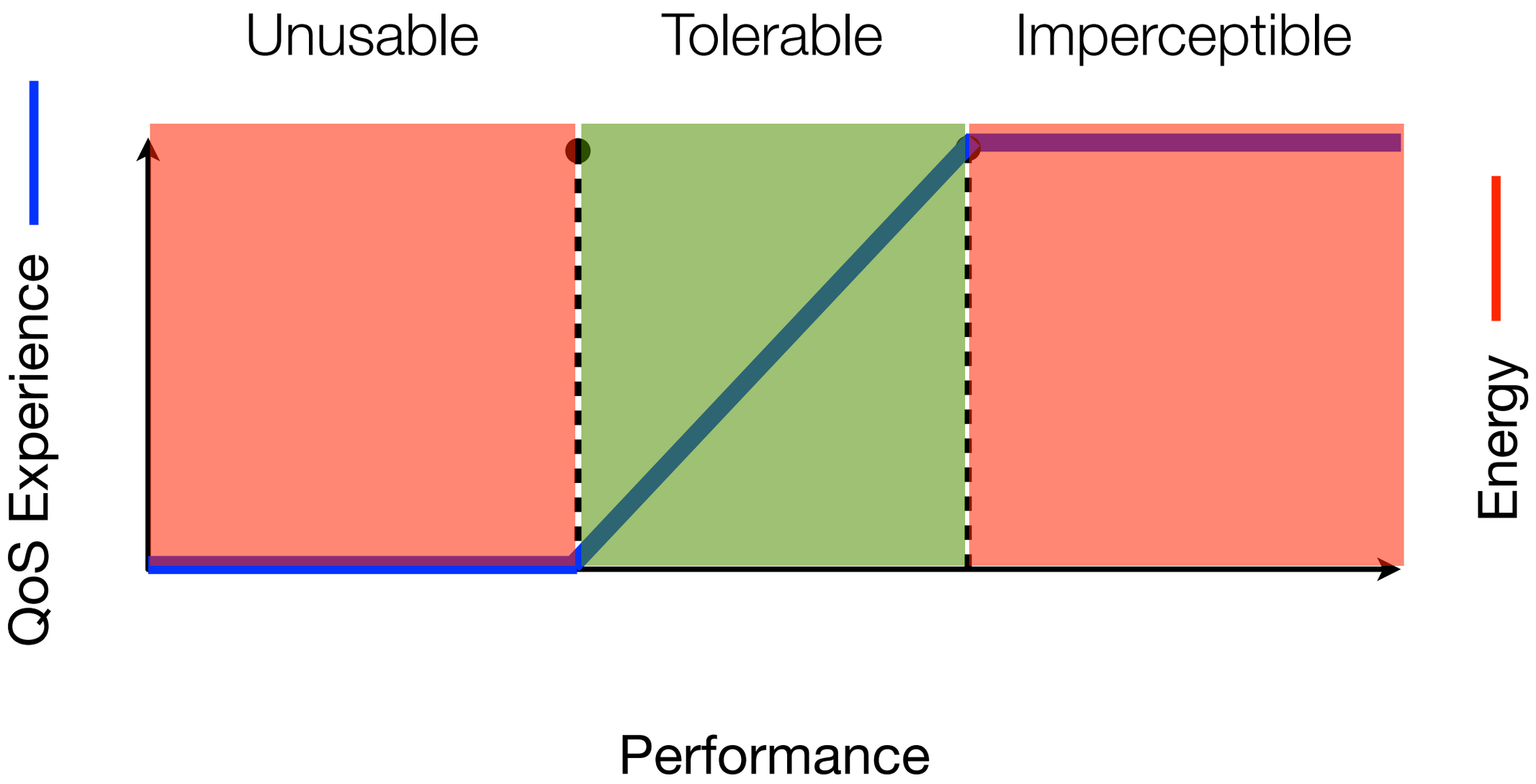
[OSDI 1996] Y. Endo et al., "Using Latency to Evaluate Interactive System Performance."

Understanding Mobile Web QoS



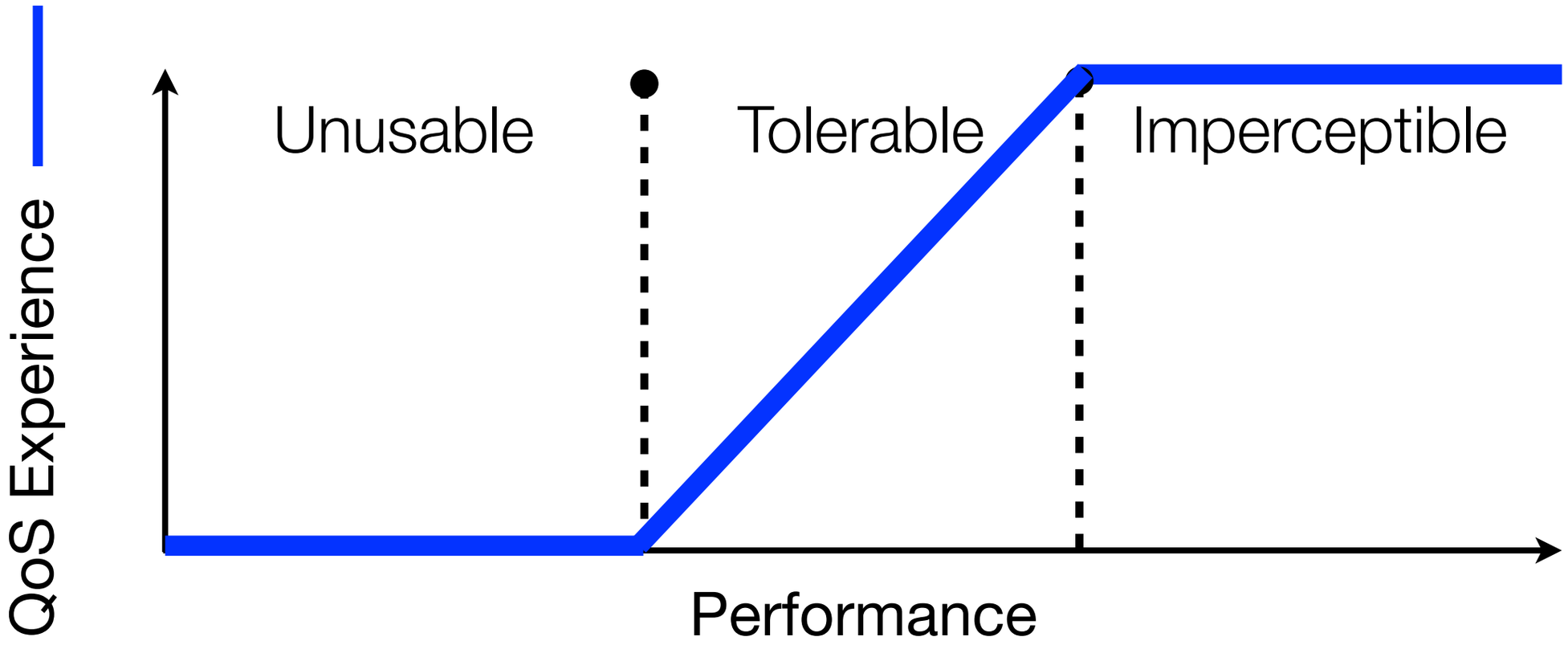
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Understanding Mobile Web QoS



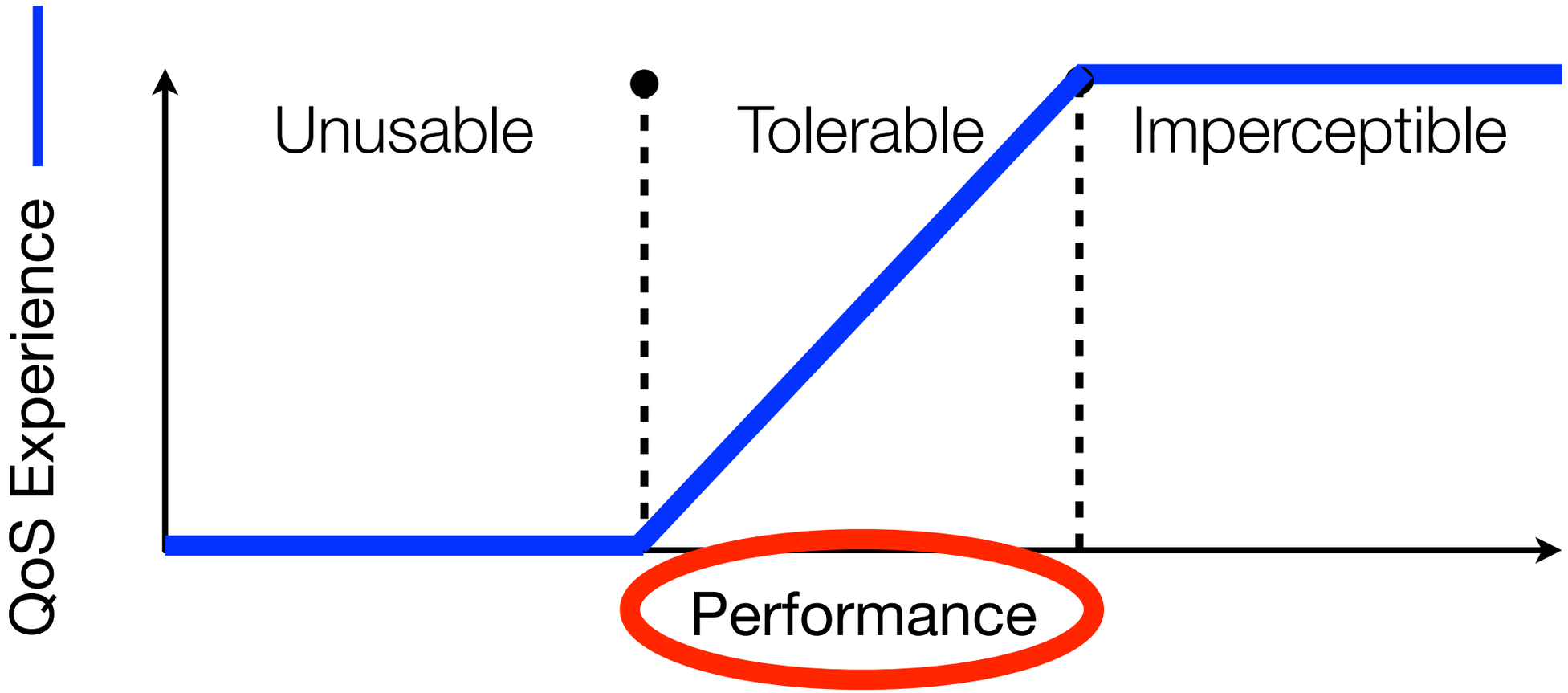
[OSDI 1996] Y. Endo et al., “Using Latency to Evaluate Interactive System Performance.”

Abstracting Mobile Web QoS



Abstracting Mobile Web QoS

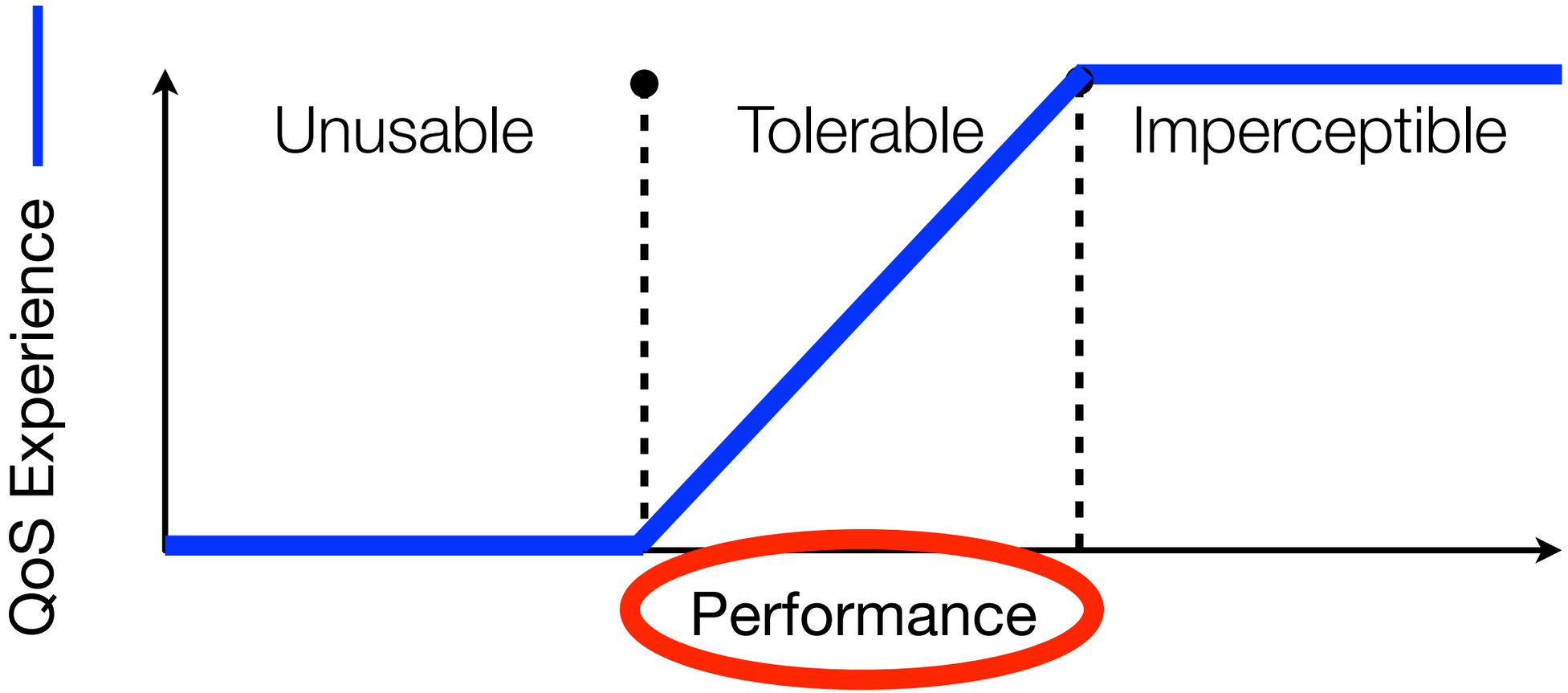
- **Performance metric**
  - Frame latency vs. Frame throughput



Abstracting Mobile Web QoS

- **Performance metric**
  - Frame latency vs. Frame throughput

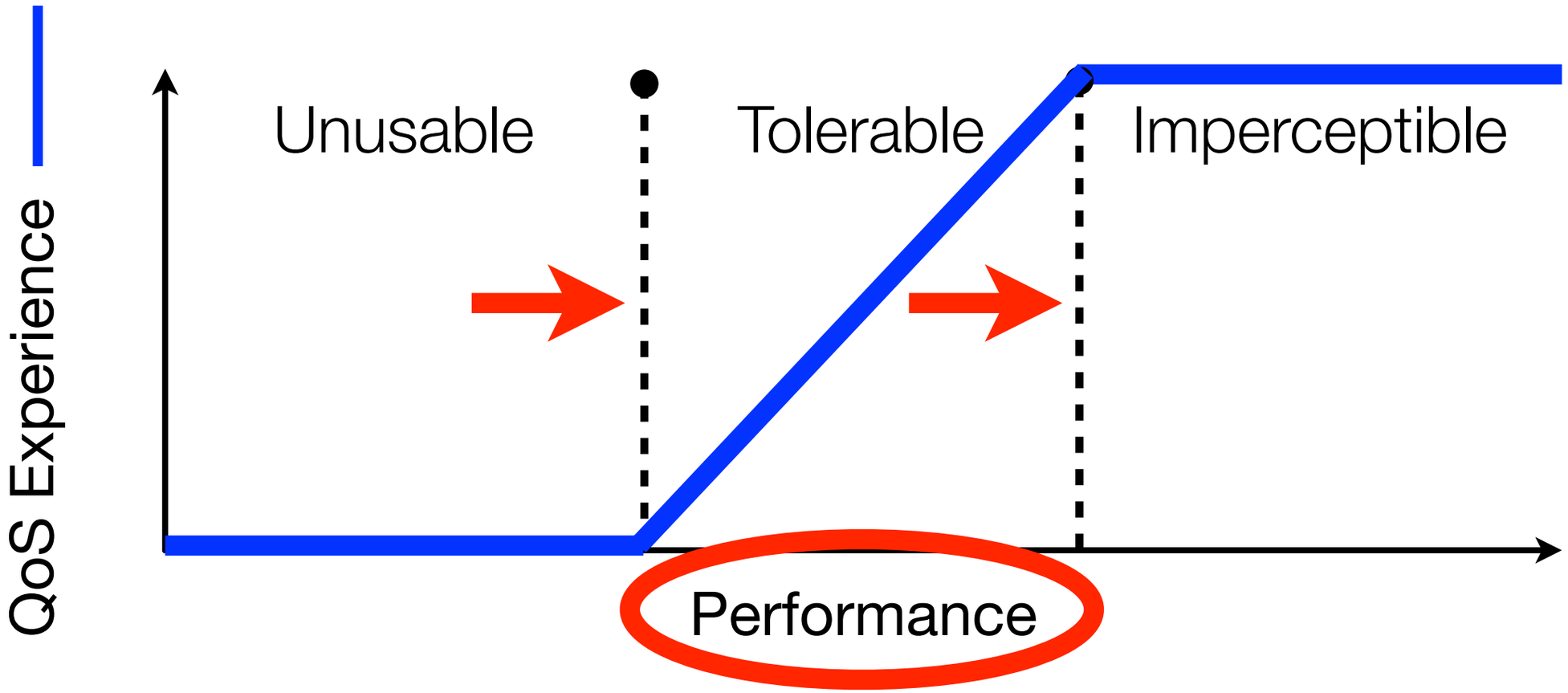
QoS Type



Abstracting Mobile Web QoS

- Performance metric
  - Frame latency vs. Frame throughput
- Threshold performance values
  - Imperceptible target vs. Usable target

QoS Type



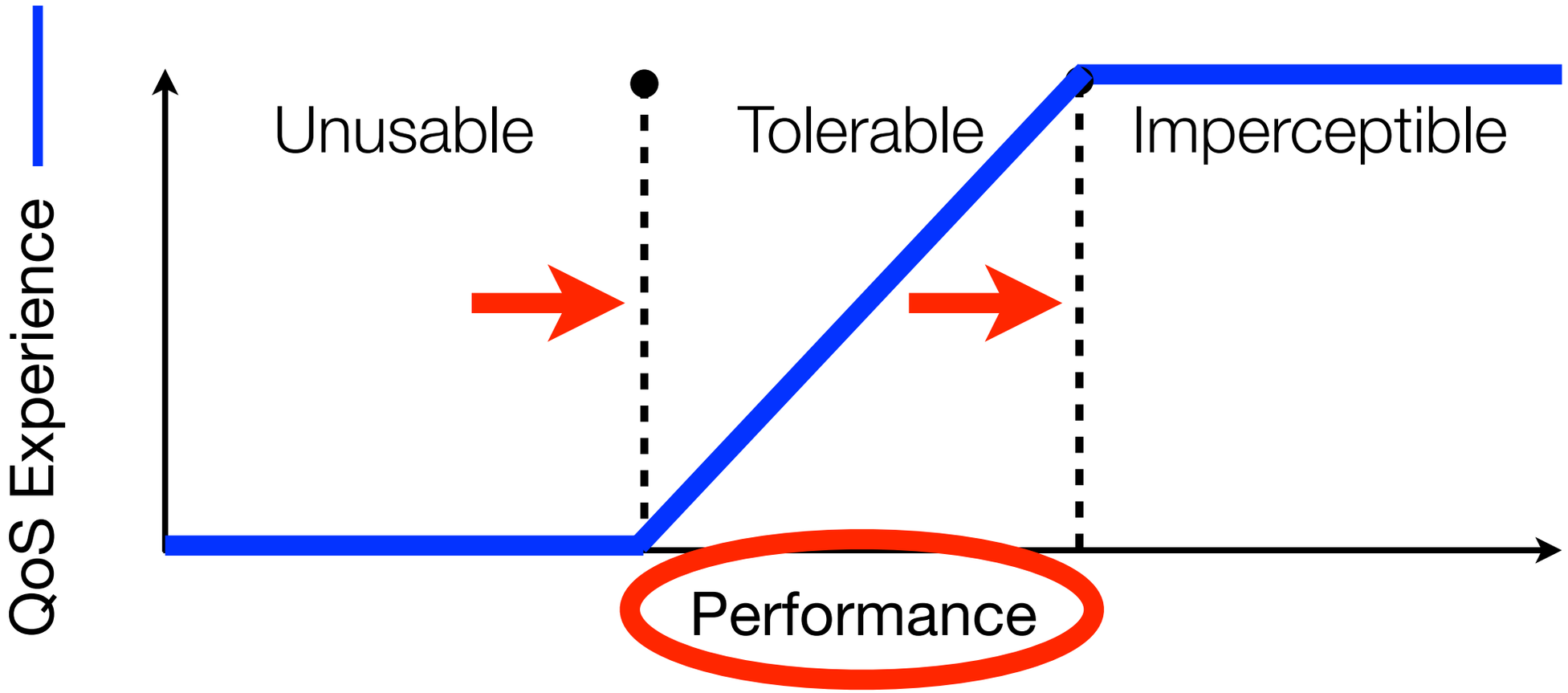
Abstracting Mobile Web QoS

- Performance metric
  - Frame latency vs. Frame throughput

QoS Type

- Threshold performance values
  - Imperceptible target vs. Usable target

QoS Target



Expressing Mobile Web QoS

Expressing Mobile Web QoS

```
<html> <head>
  <script>
    function animateMove() {
      /* Animation code omitted */
    }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!– other elements -->
</body> </html>
```

Expressing Mobile Web QoS

element

```
<html> <head>
  <script>
    function animateMove() {
      /* Animation code omitted */
    }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!– other elements -->
</body> </html>
```

Expressing Mobile Web QoS

element **event**

```
<html> <head>
  <script>
    function animateMove() {
      /* Animation code omitted */
    }
  </script> </head> <body>
  <div ontouchend=“animateMove()”>
  <div/> <!– other elements -->
</body> </html>
```

Expressing Mobile Web QoS

element **event**

```
<html> <head>
  <script>
    function animateMove() {
      /* Animation code omitted */
    }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!– other elements -->
</body> </html>
```

Expressing Mobile Web QoS

element **event**

```
<html> <head>
  }
</script> </head> <body>
<div ontouchend="animateMove()">
<div/> <!– other elements -->
</body> </html>
```

Expressing QoS at an event granularity

Expressing Mobile Web QoS

Annotation

```
                element   event

<html> <head>
  <style>


  </style>
  <script>
   function animateMove() {
     /* Animation code omitted */
   }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!— other elements -->
</body> </html>
```

Expressing Mobile Web QoS

Annotation

```
                element   event

<html> <head>
  <style>
    div {
      ontouchend
    }
  </style>
  <script>
   function animateMove() {
     /* Animation code omitted */
   }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!— other elements -->
</body> </html>
```

Expressing Mobile Web QoS

Annotation

```
element {event: Type, Target}
```

```
<html> <head>
  <style>
    div {
      ontouchend: throughput, low;
    }
  </style>
  <script>
   function animateMove() {
     /* Animation code omitted */
   }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!– other elements -->
</body> </html>
```

Expressing Mobile Web QoS


```
element {event: Type, Target}
```

Annotation

```
<html> <head>
  <style>
    div {
      ontouchend: throughput, low;
    }
  </style>
  <script>
   function animateMove() {
     /* Animation code omitted */
   }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!– other elements -->
</body> </html>
```

Expressing Mobile Web QoS


```
element {event: Type, Target}
```

Annotation

```
<html> <head>
  <style>
    div {
      ontouchend: throughput, low;
    }
  </style>
  <script>
    function animateMove() {
      /* Animation code omitted */
    }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!– other elements -->
</body> </html>
```

Expressing Mobile Web QoS


```
element {event: Type, Target}
```

Annotation

```
<html> <head>
  <style>
    div {
      ontouchend: throughput, low;
    }
  </style>
  <script>
    function newAnimateMove() {
      /* New animation code */
    }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!— other elements -->
</body> </html>
```

Expressing Mobile Web QoS


```
element {event: Type, Target}
```

Annotation

```
<html> <head>
  <style>
    div {
      ontouchend: throughput, low;
    }
  </style>
  <script>
    function newAnimateMove() {
      /* New animation code */
    }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!— other elements -->
</body> </html>
```

Implementation independent

Expressing Mobile Web QoS


```
element {event: Type, Target}
```

Annotation

```
<html> <head>
  <style>
    div {
      ontouchend: throughput, low;
    }
  </style>
  <script>
    function newAnimateMove() {
      /* New animation code */
    }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!– other elements -->
</body> </html>
```

Implementation independent

Expressing Mobile Web QoS

CSS
Cascading Style Sheet

```
element {event: Type, Target}
```

Annotation

```
<html> <head>
  <style>
    div {
      ontouchend: throughput, low;
    }
  </style>
  <script>
    function newAnimateMove() {
      /* New animation code */
    }
  </script> </head> <body>
  <div ontouchend="animateMove()">
  <div/> <!– other elements -->
</body> </html>
```

Implementation independent

Non-interfering w.r.t. functionality

GreenWeb Annotation Process

Original application

Manual Annotation

GreenWeb-annotated application

GreenWeb Annotation Process

Original application

Automatic Annotation?

GreenWeb-annotated application

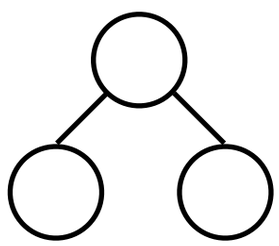
GreenWeb Annotation Process

- **AutoGreen**: *automatically* reasons about and inserts **GreenWeb** annotations

GreenWeb Annotation Process

- **AutoGreen**: *automatically* reasons about and inserts **GreenWeb** annotations

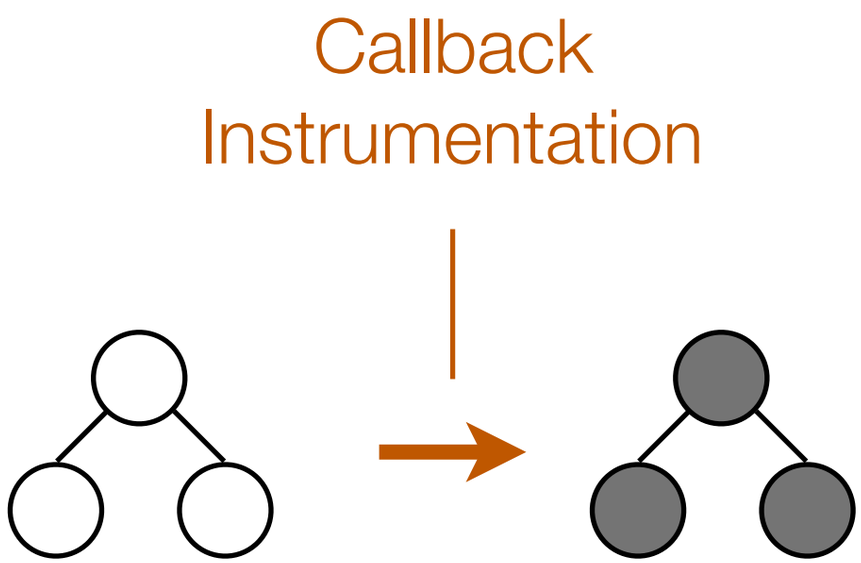
GreenWeb Annotation Process



GreenWeb-annotated application

- **AutoGreen**: *automatically* reasons about and inserts **GreenWeb** annotations

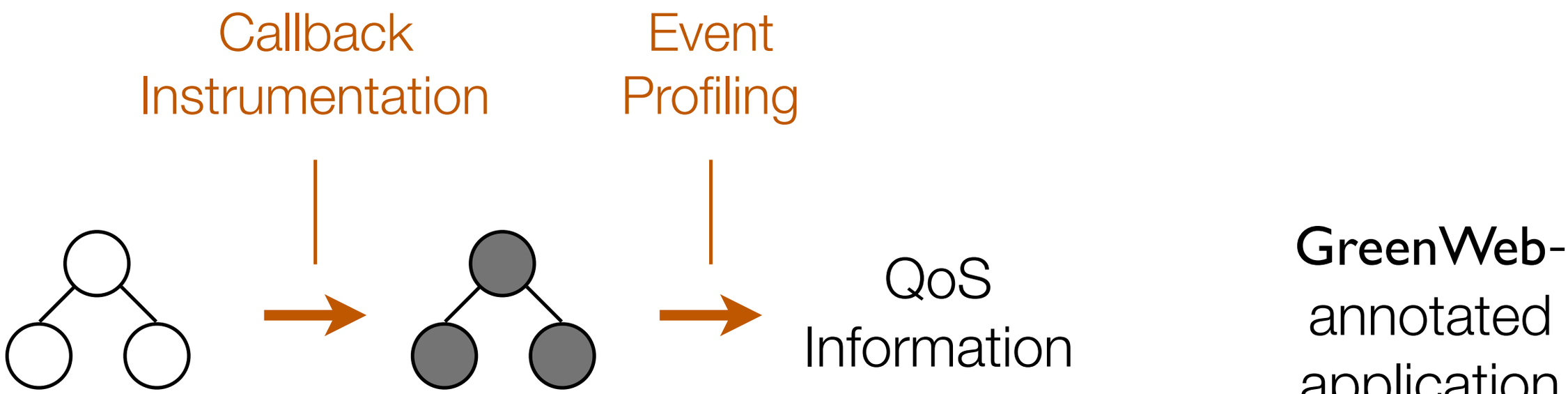
GreenWeb Annotation Process



GreenWeb-annotated application

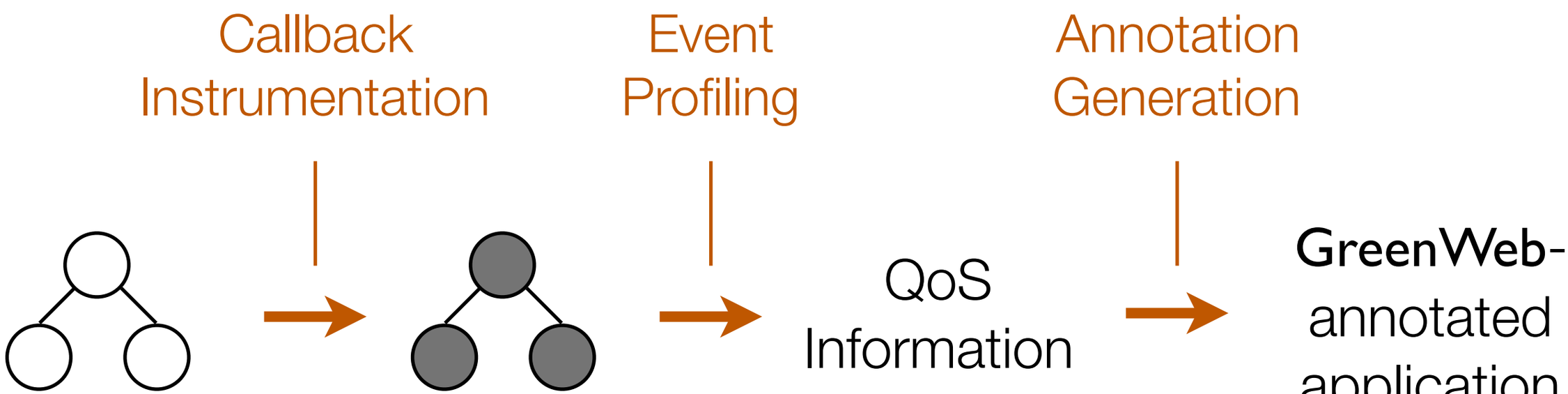
- **AutoGreen**: *automatically* reasons about and inserts **GreenWeb** annotations

GreenWeb Annotation Process



- **AutoGreen**: *automatically* reasons about and inserts **GreenWeb** annotations

GreenWeb Annotation Process



- **AutoGreen**: *automatically* reasons about and inserts **GreenWeb** annotations

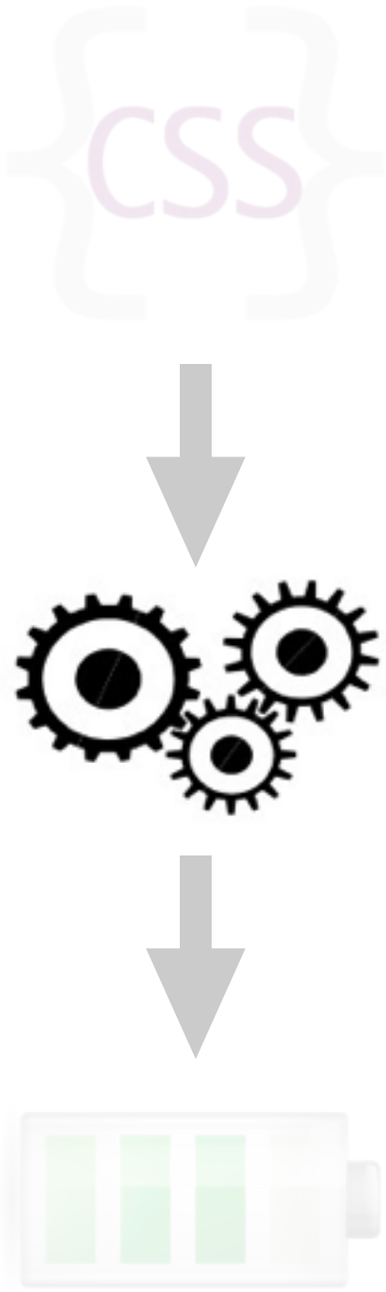
GreenWeb: Language for Energy-Efficiency

- **Language abstractions** for expressing QoS

- Runtime
  the QoS constraints

- Result
  hardware/software implementations

GreenWeb: Language for Energy-Efficiency



- Language abstractions
- **Runtime** that saves energy while meeting the QoS constraints
- Result
  hardware/software implementations

GreenWeb Runtime Overview

GreenWeb Runtime Overview

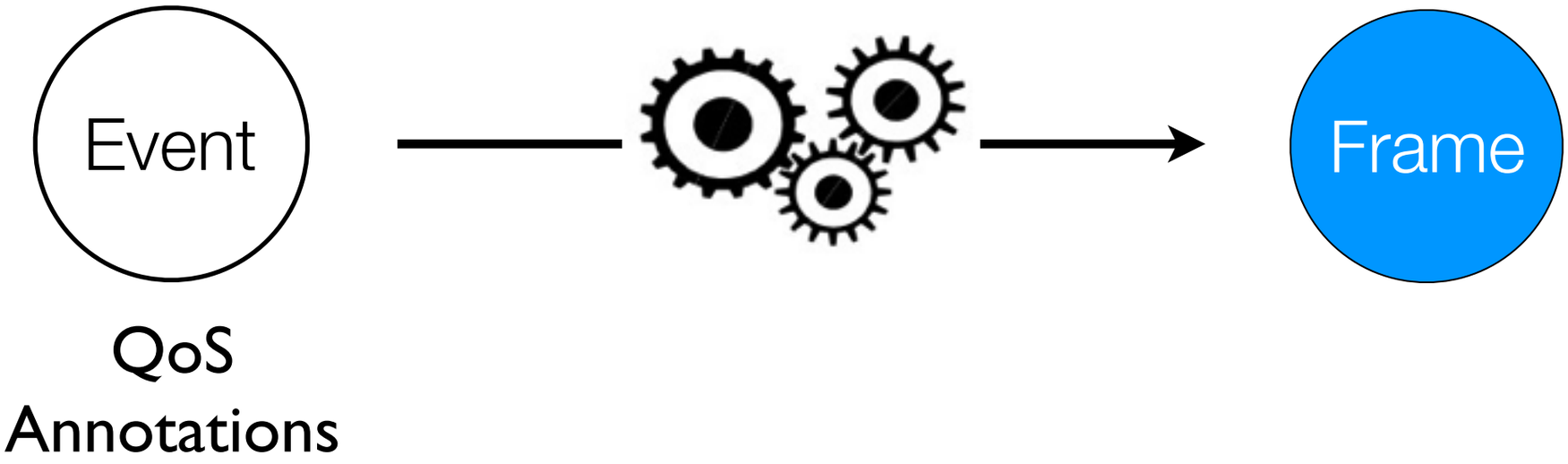
GreenWeb Runtime Overview

QoS
Annotations

GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently



GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently

QoS type: latency

QoS target: 16 ms

GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently

QoS type: latency

QoS target: 16 ms

GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently

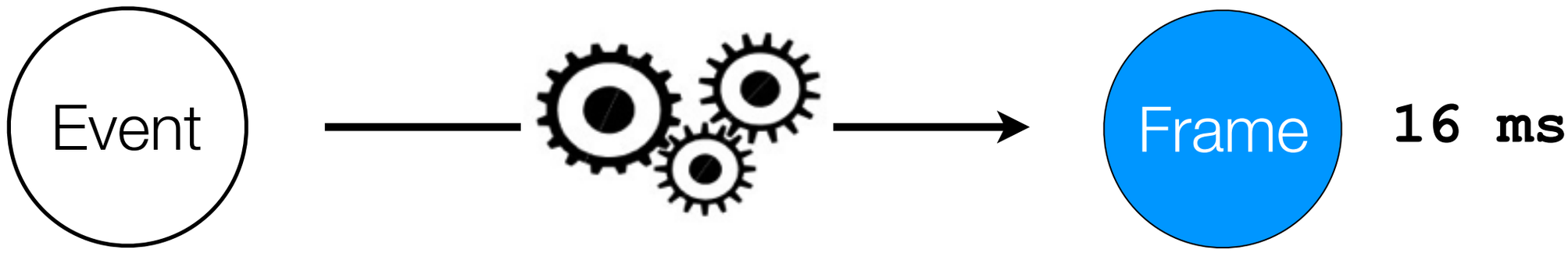
QoS type: latency

QoS target: 16 ms

GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently



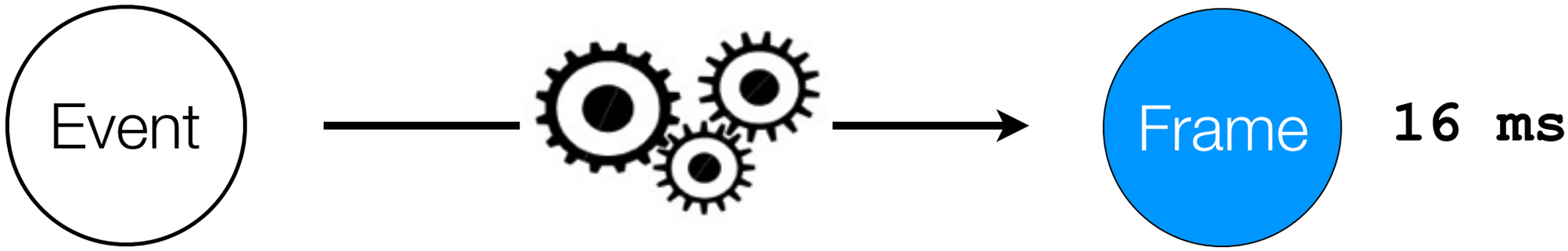
QoS type: latency

QoS target: 16 ms

GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently



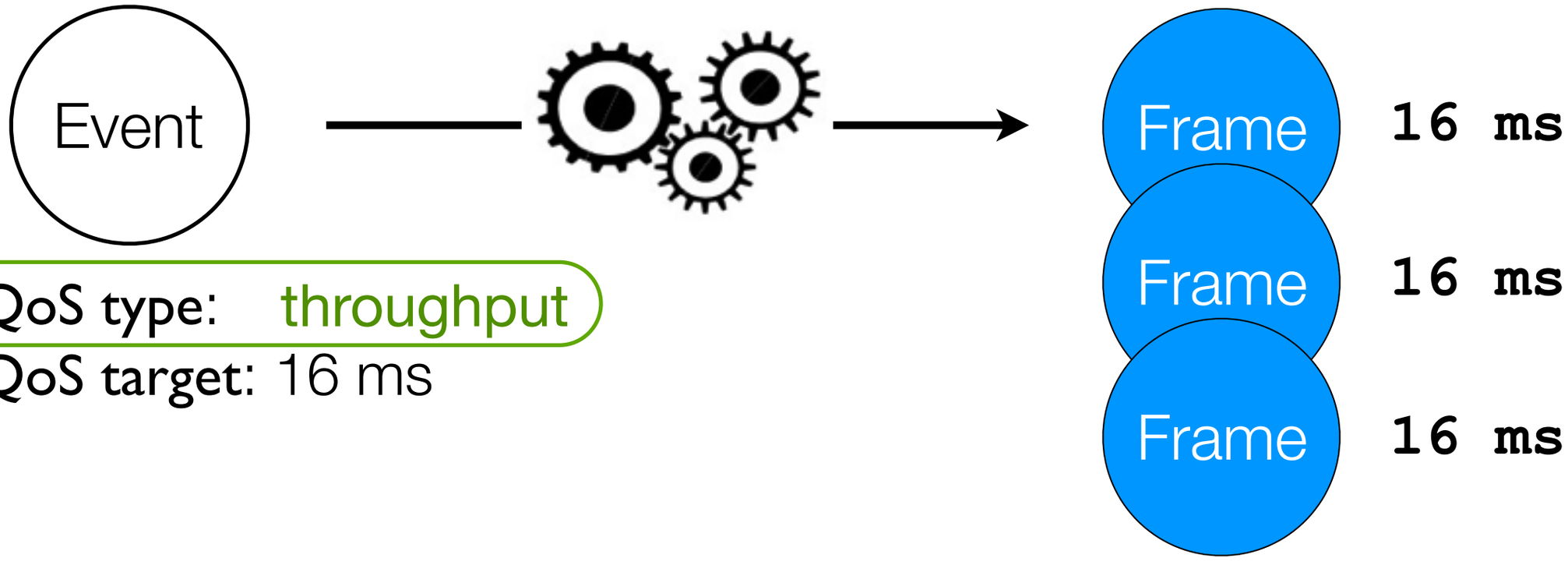
QoS type: throughput

QoS target: 16 ms

GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently



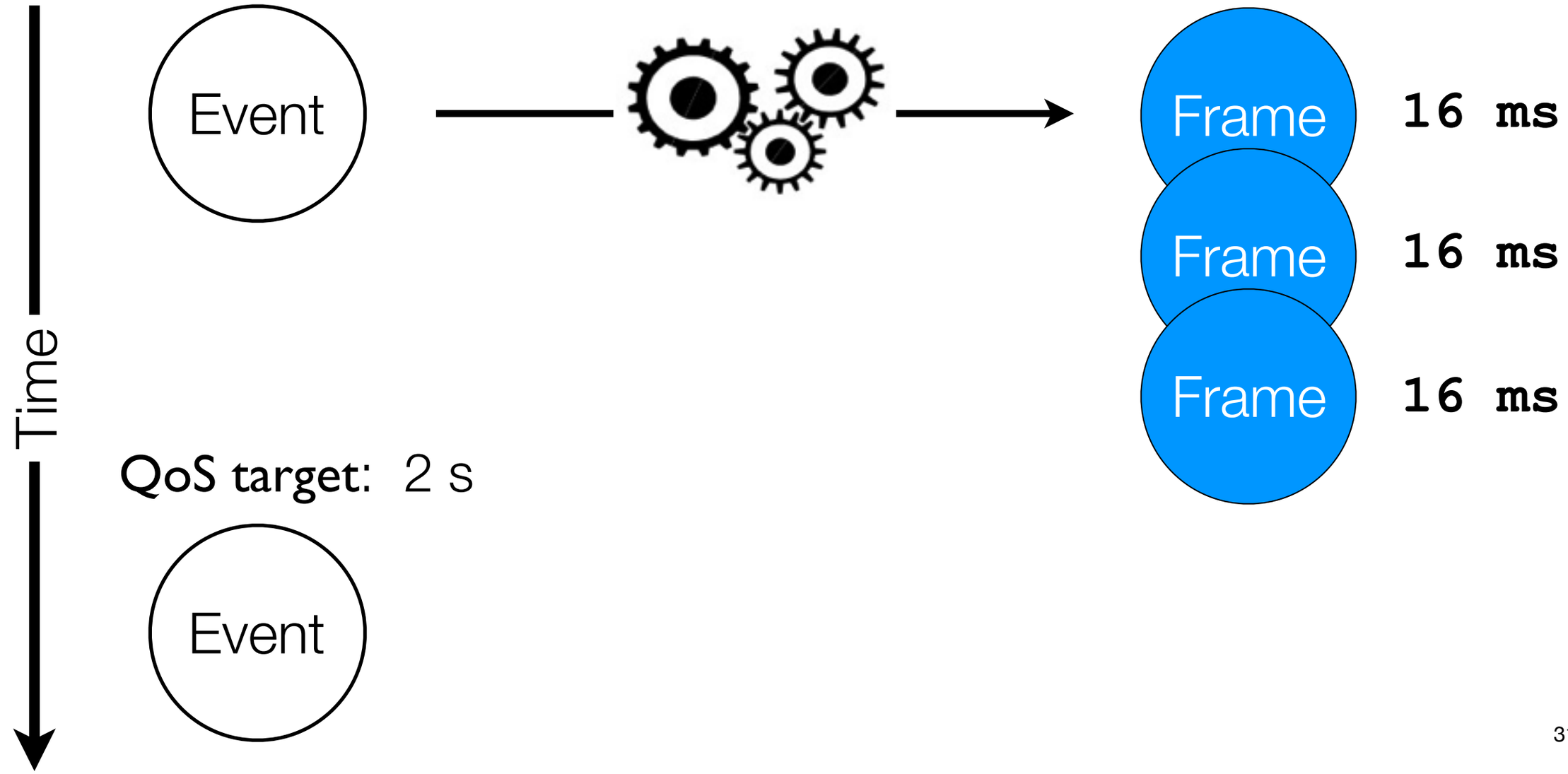
QoS type: throughput

QoS target: 16 ms

GreenWeb Runtime Overview

Runtime Objective

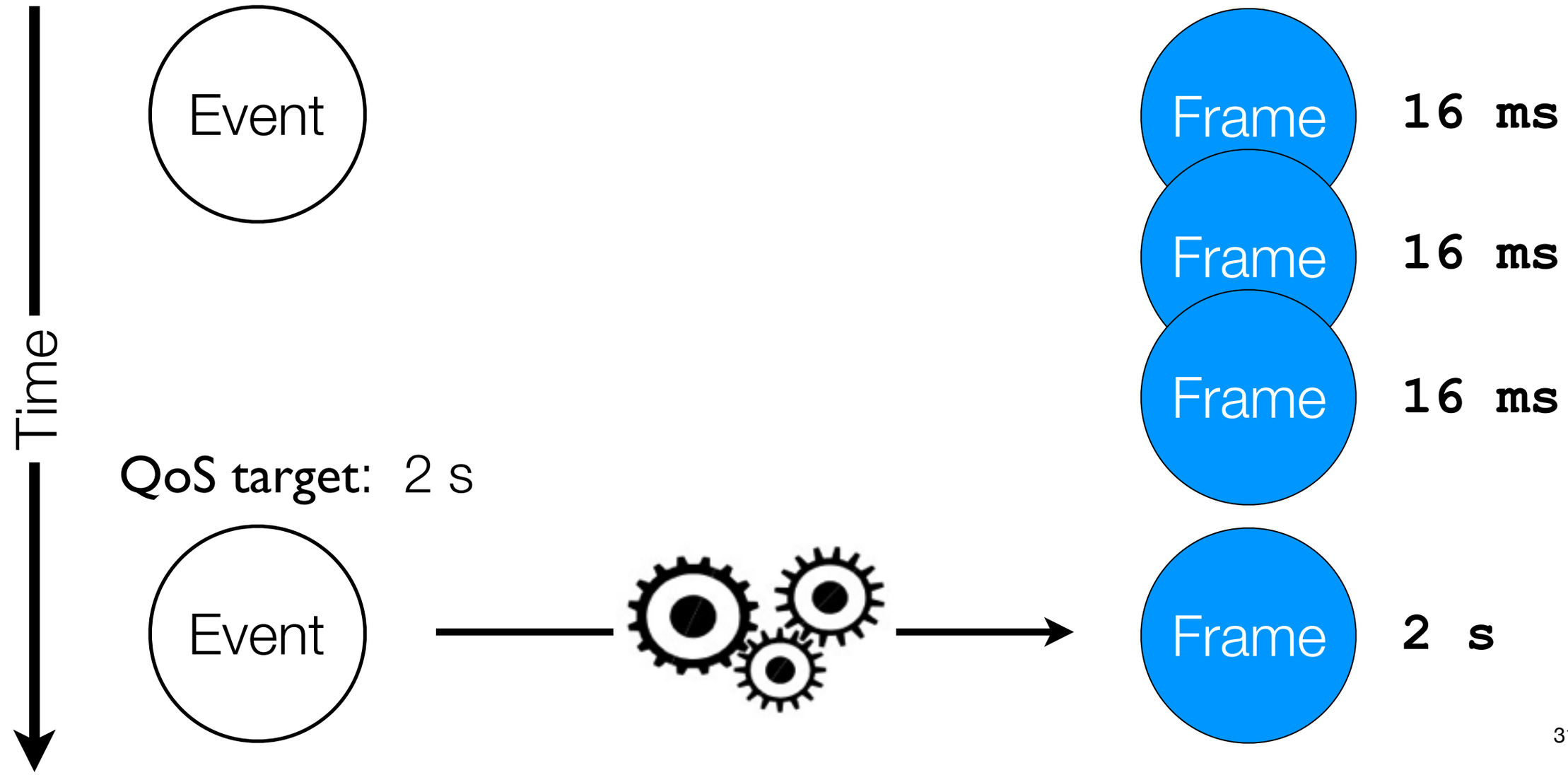
Enforcing event-level QoS at the frame-level energy-efficiently



GreenWeb Runtime Overview

Runtime Objective

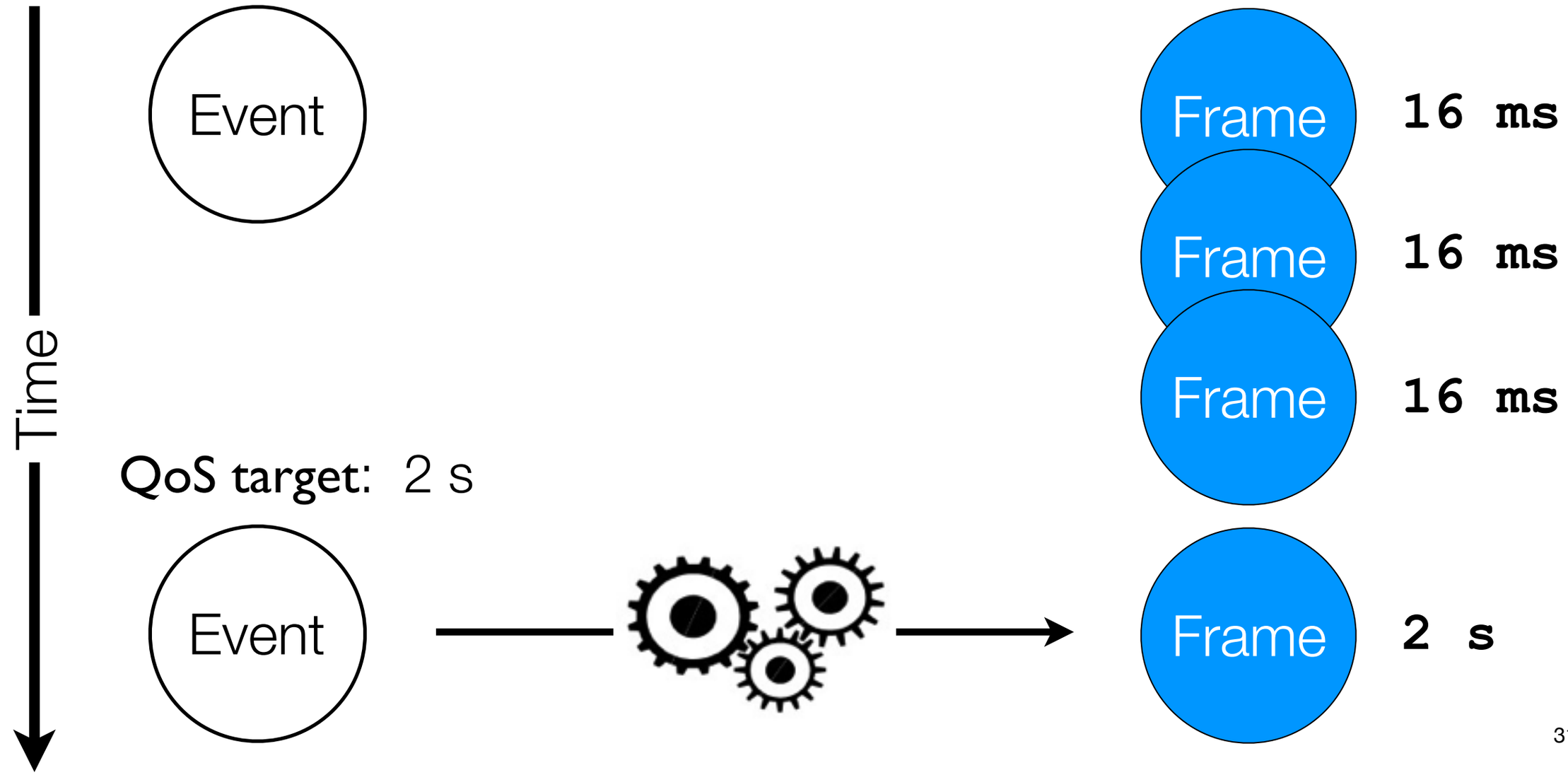
Enforcing event-level QoS at the frame-level energy-efficiently



GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently {



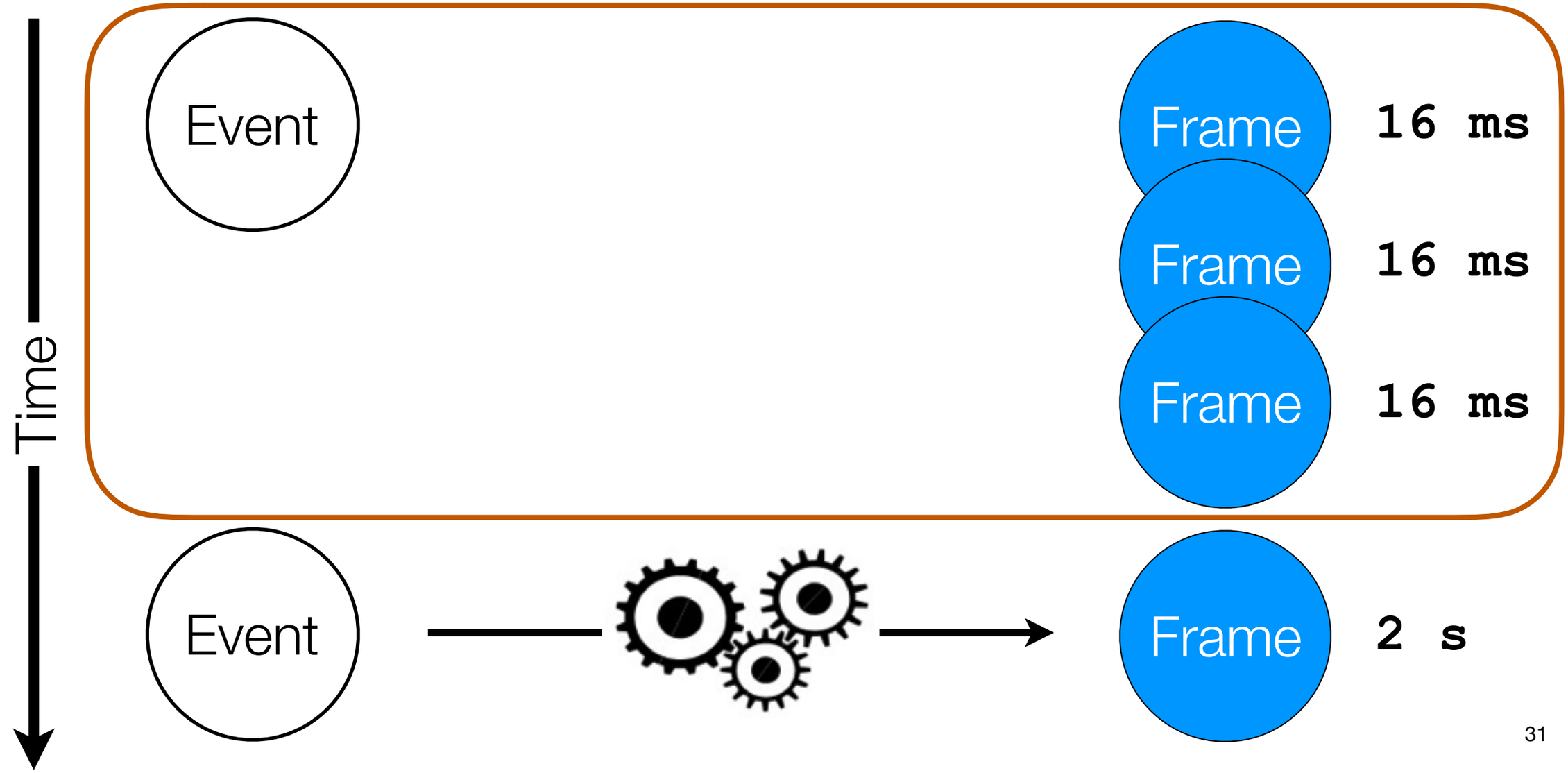
GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently

{

1. Frame Association

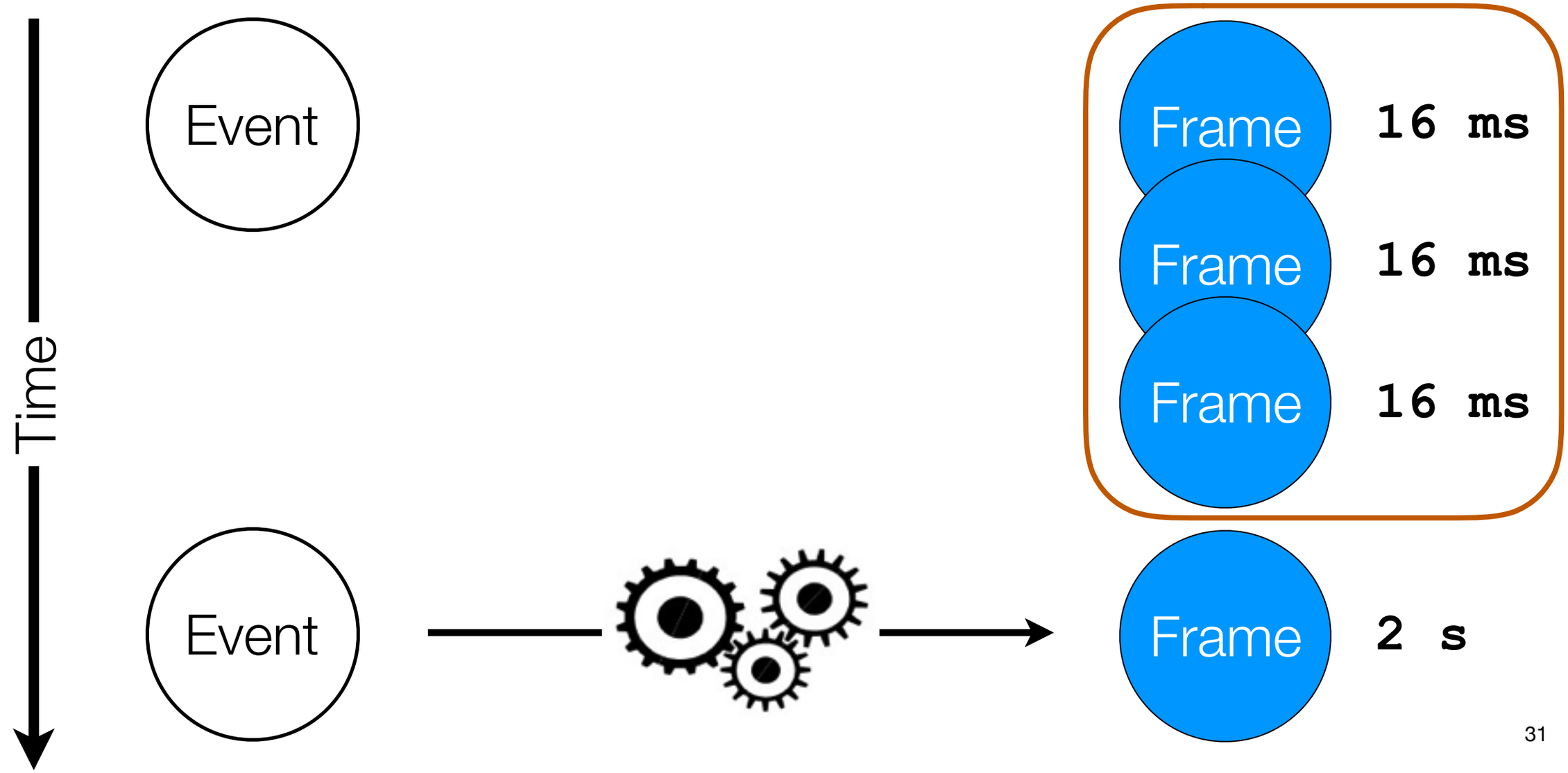


GreenWeb Runtime Overview

Runtime Objective

Enforcing event-level QoS at the frame-level energy-efficiently

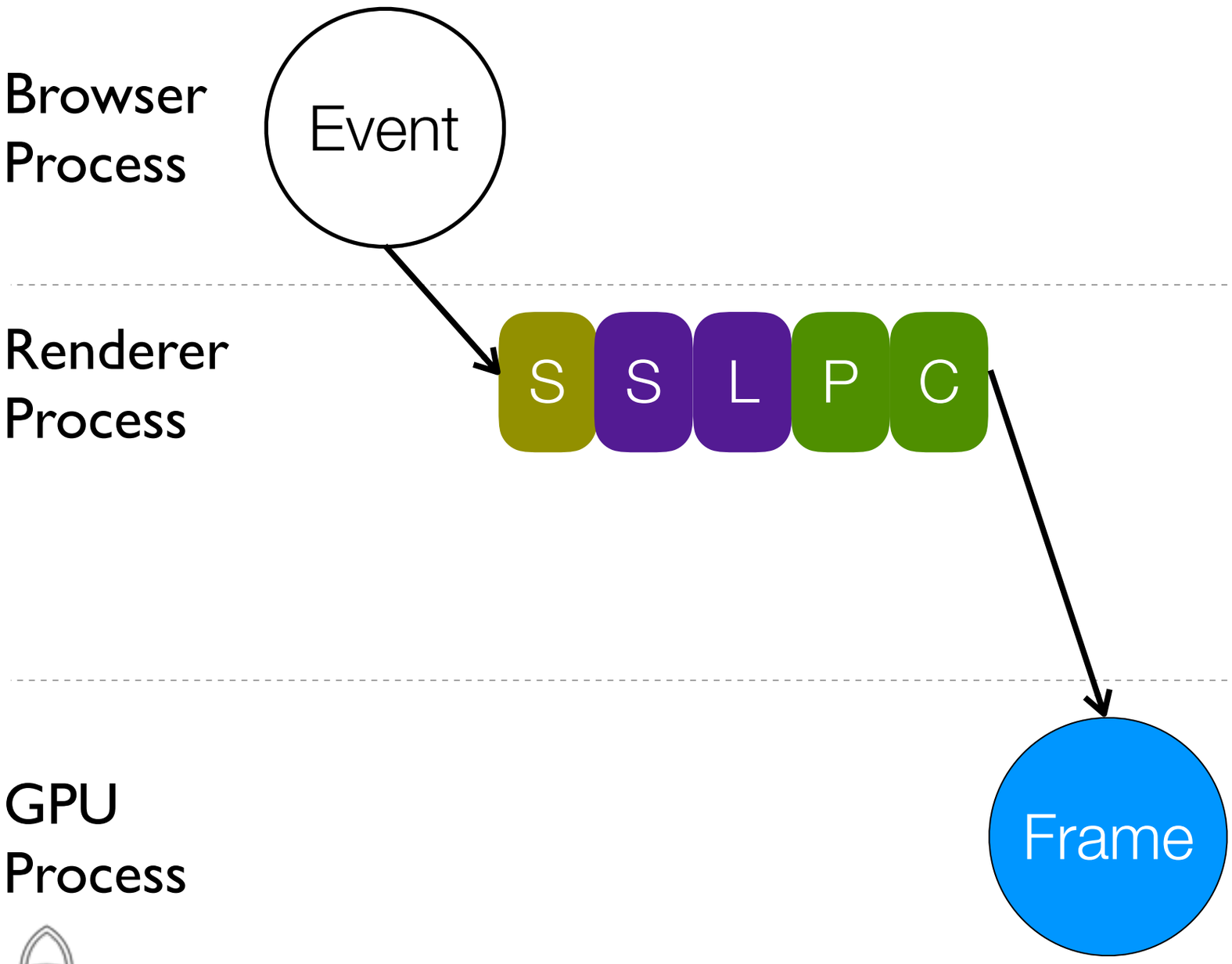
1. Frame Association
2. Frame Scheduling



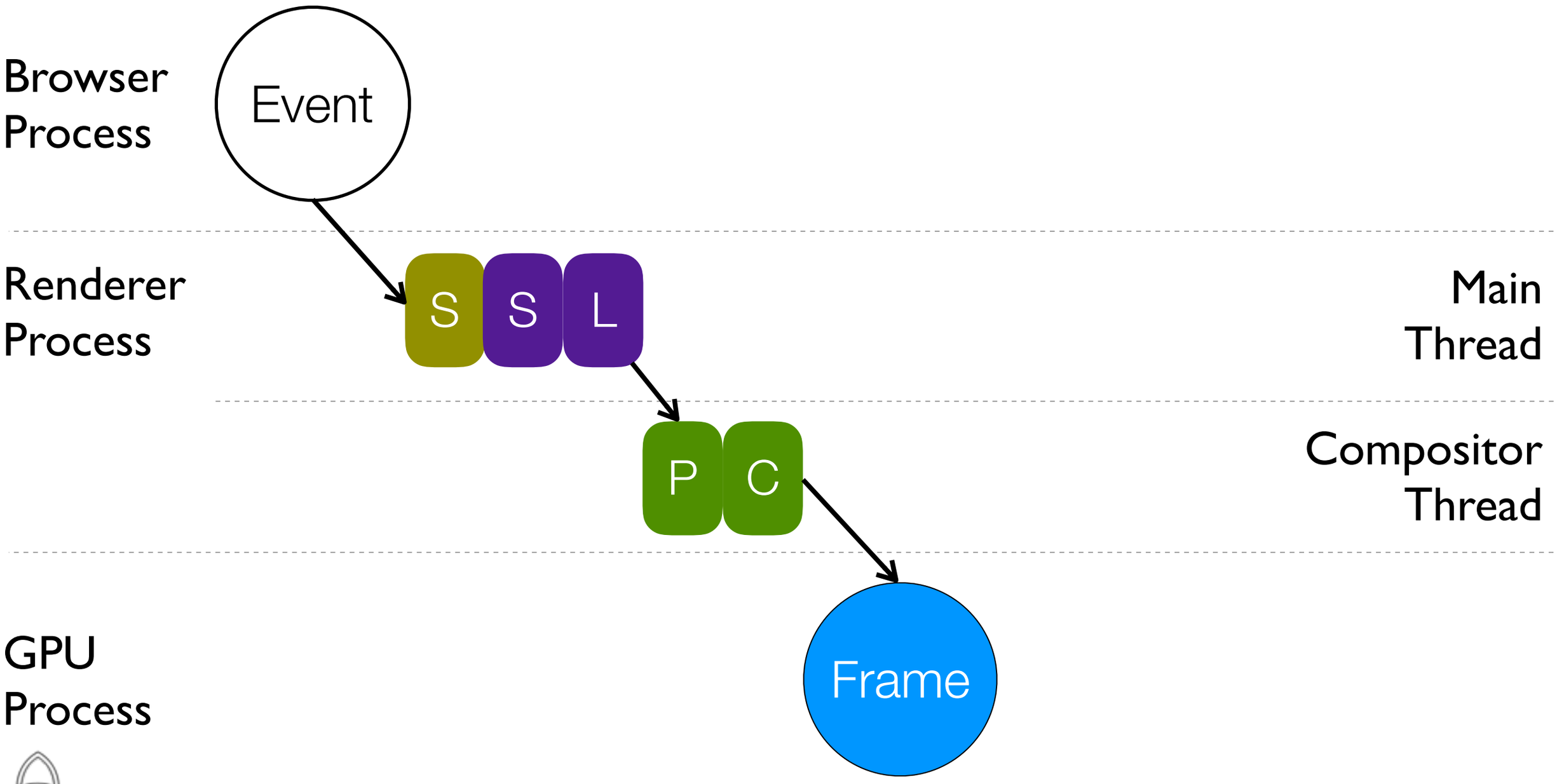
Frame Association

Frame Association

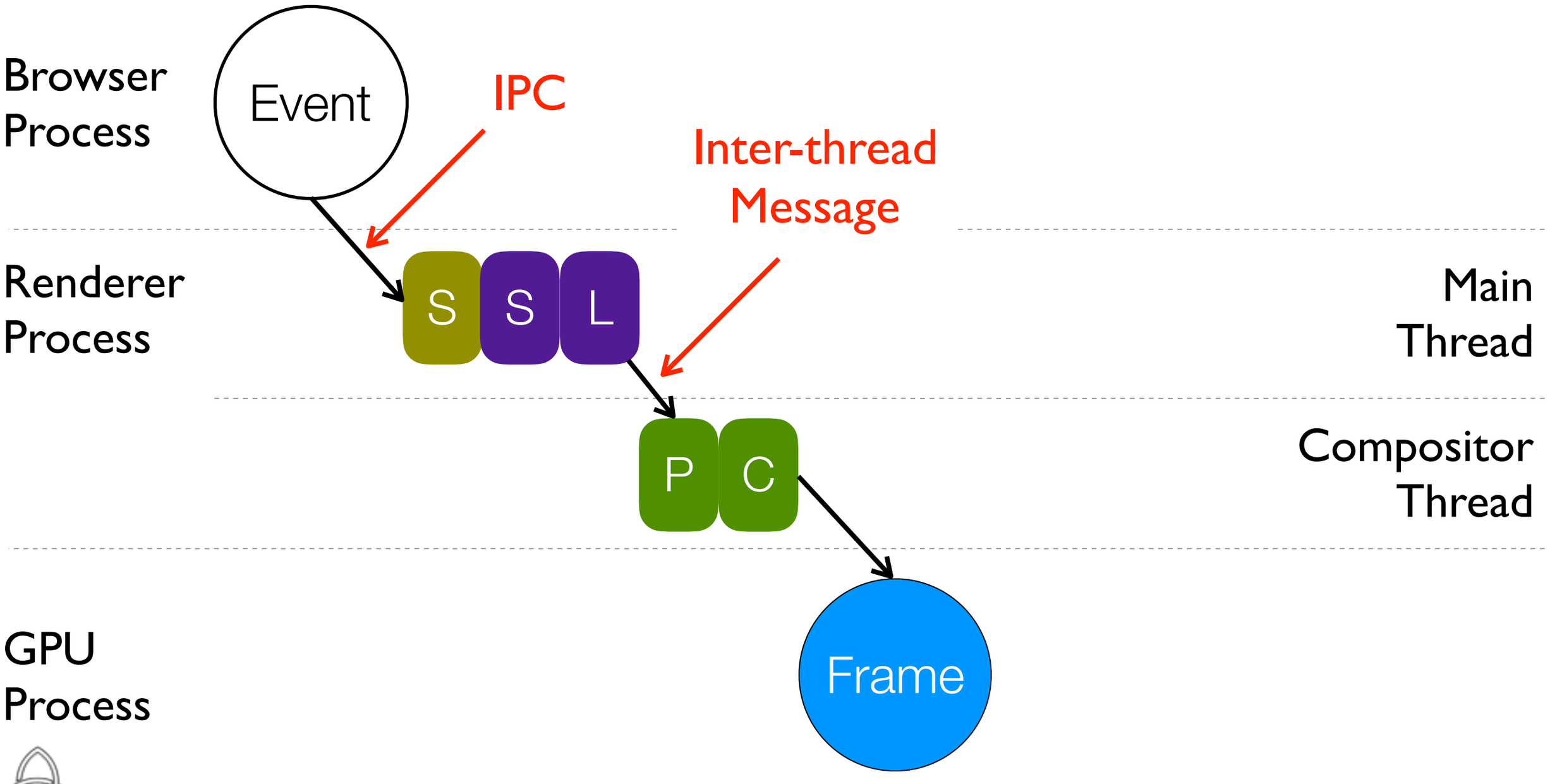
Frame Association



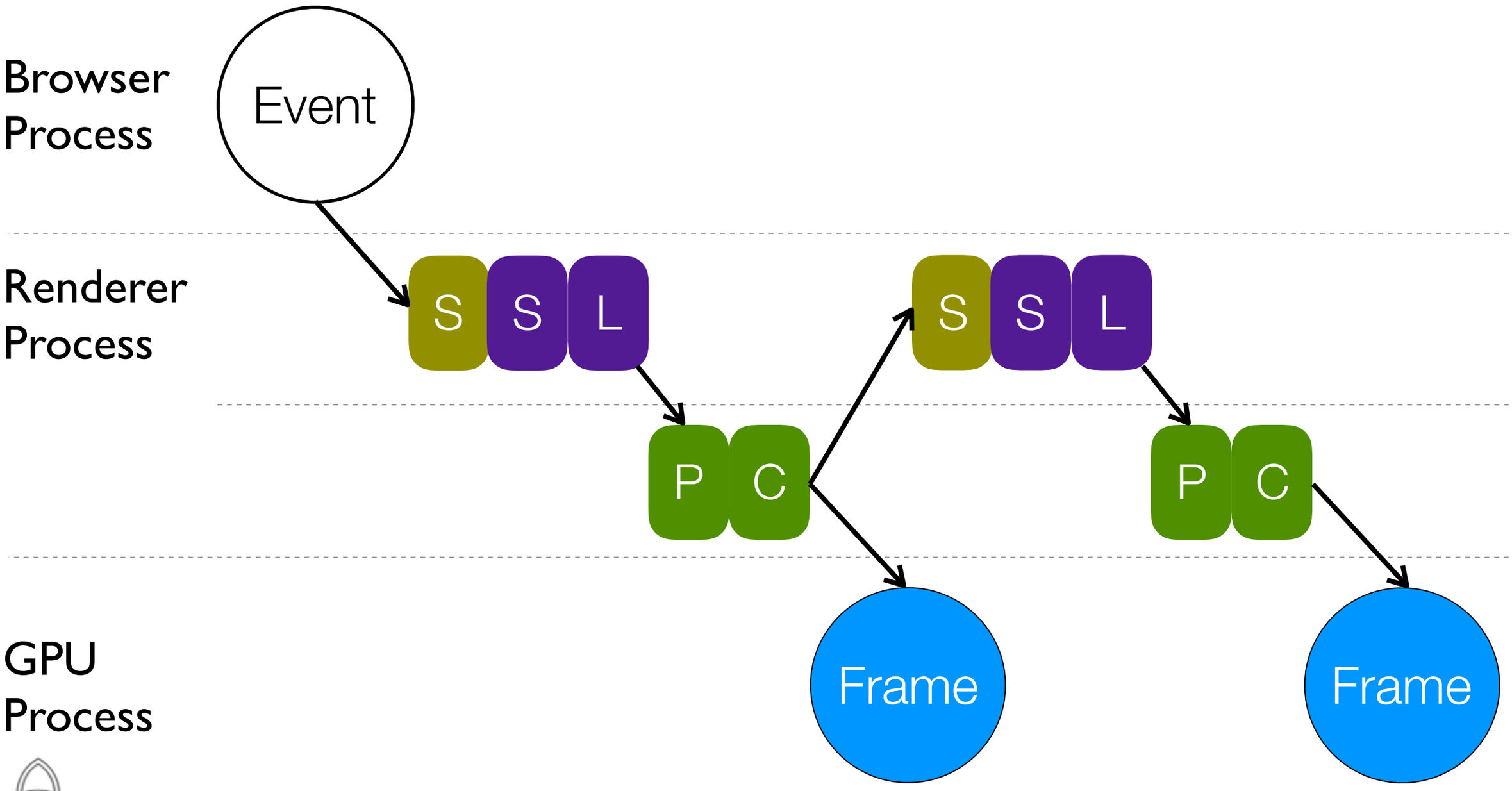
Frame Association



Frame Association

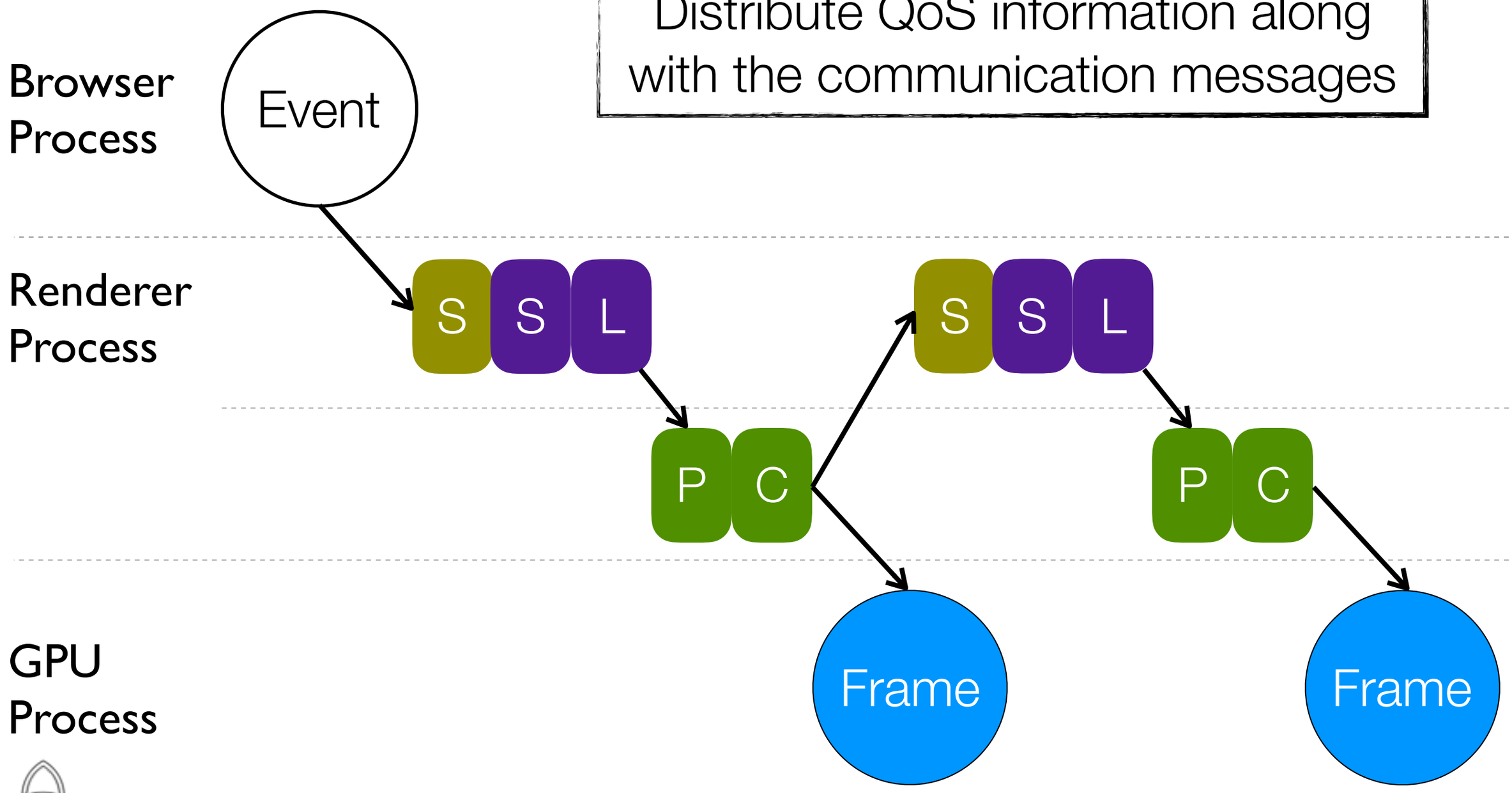


Frame Association



Frame Association

Distribute QoS information along with the communication messages



Choices of Energy-saving Techniques

GreenWeb can support a range of energy saving techniques

Choices of Energy-saving Techniques

GreenWeb can support a range of energy saving techniques

- Dynamic resolution scaling [MobiCom 2015]
- Power-saving display colors [MobiSys 2012]
- Selective resource loading [NSDI 2015]

Choices of Energy-saving Techniques

GreenWeb can support a range of energy saving techniques

- Dynamic resolution scaling [MobiCom 2015]
- Power-saving display colors [MobiSys 2012]
- Selective resource loading [NSDI 2015]
- ACMP-based hardware mechanism (WebRT)

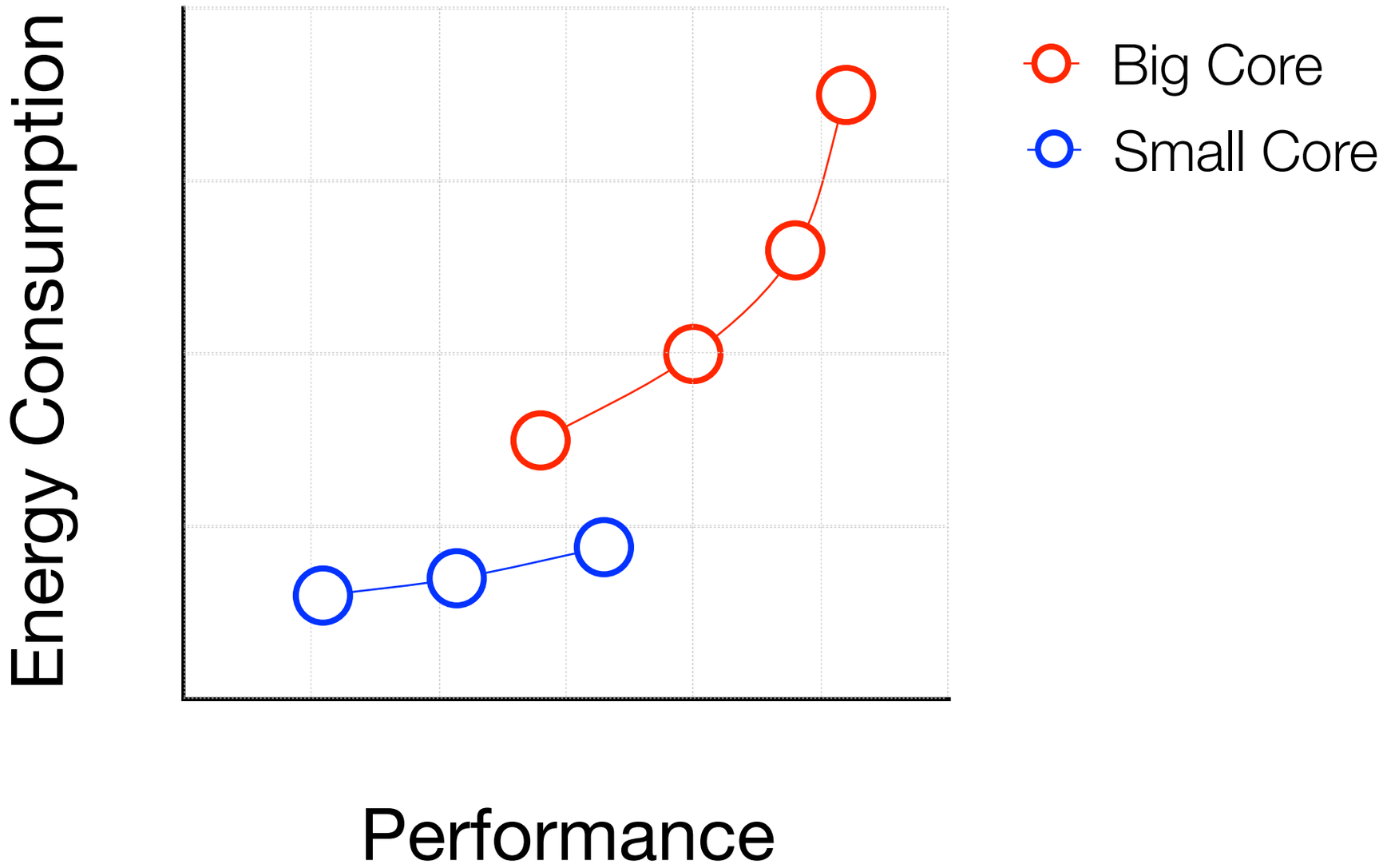
Asymmetric Chip-multiprocessor

Asymmetric Chip-multiprocessor

- Offer a large performance-energy trade-off space

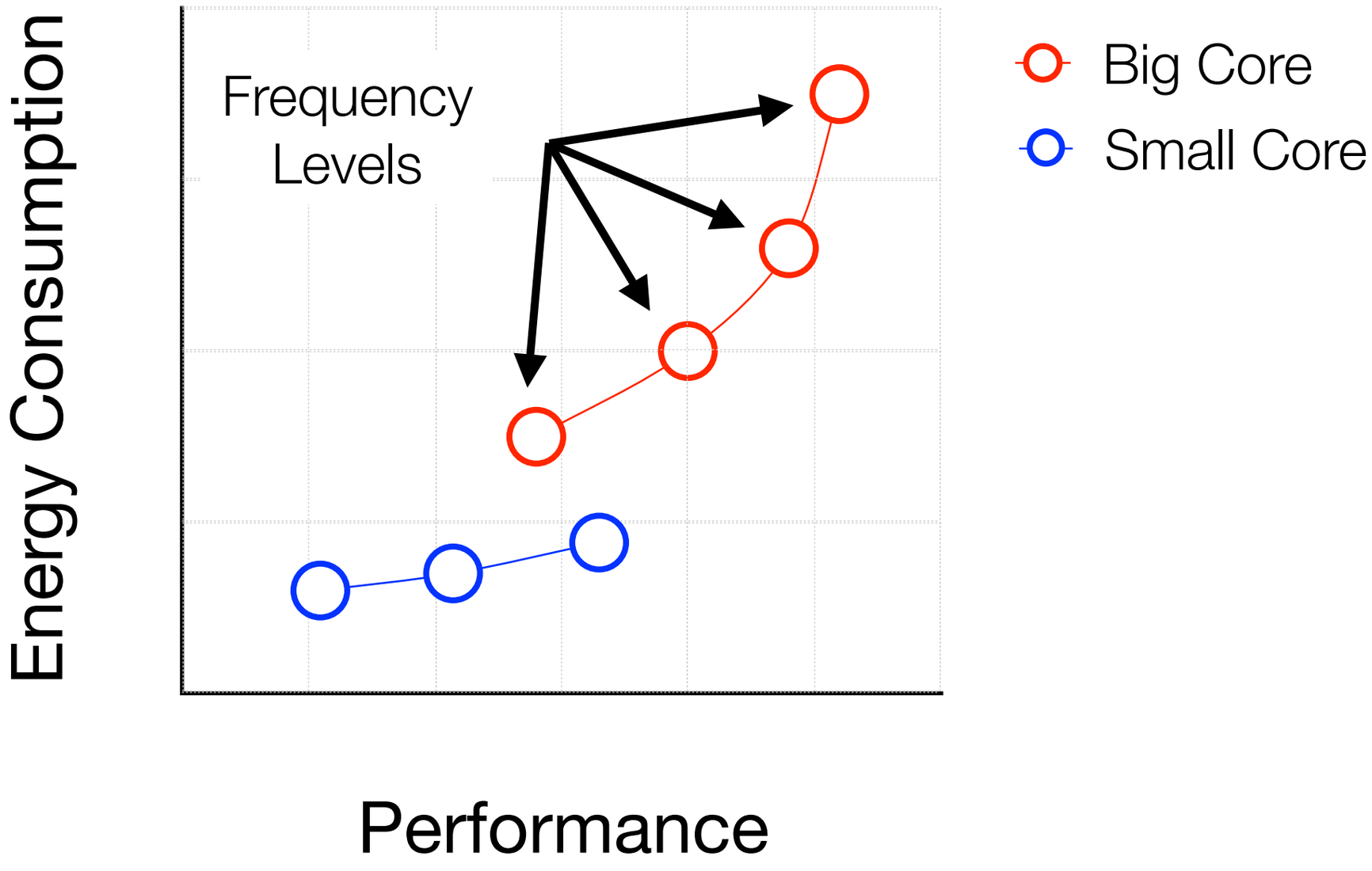
Asymmetric Chip-multiprocessor

- Offer a large performance-energy trade-off space



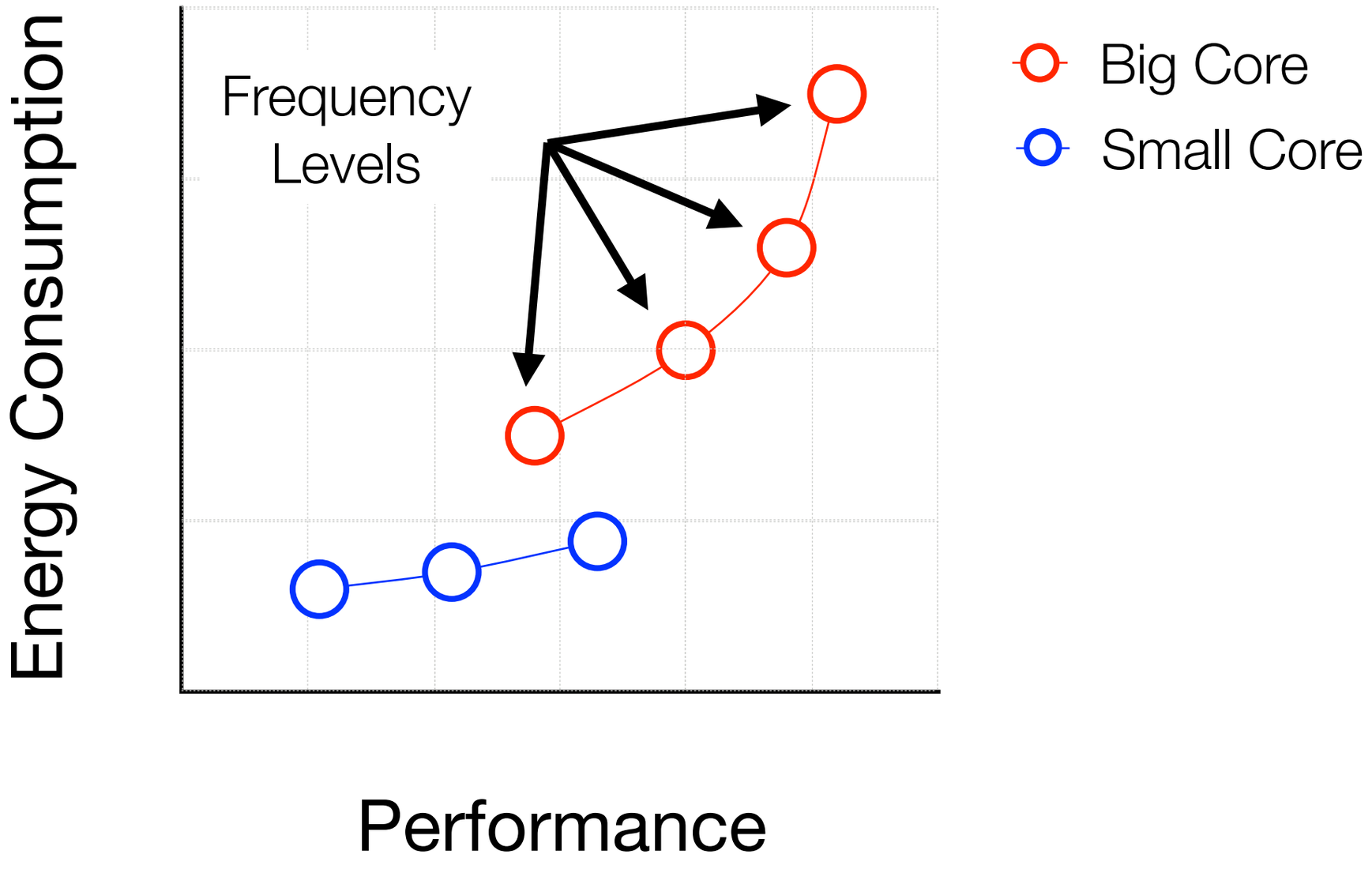
Asymmetric Chip-multiprocessor

- Offer a large performance-energy trade-off space



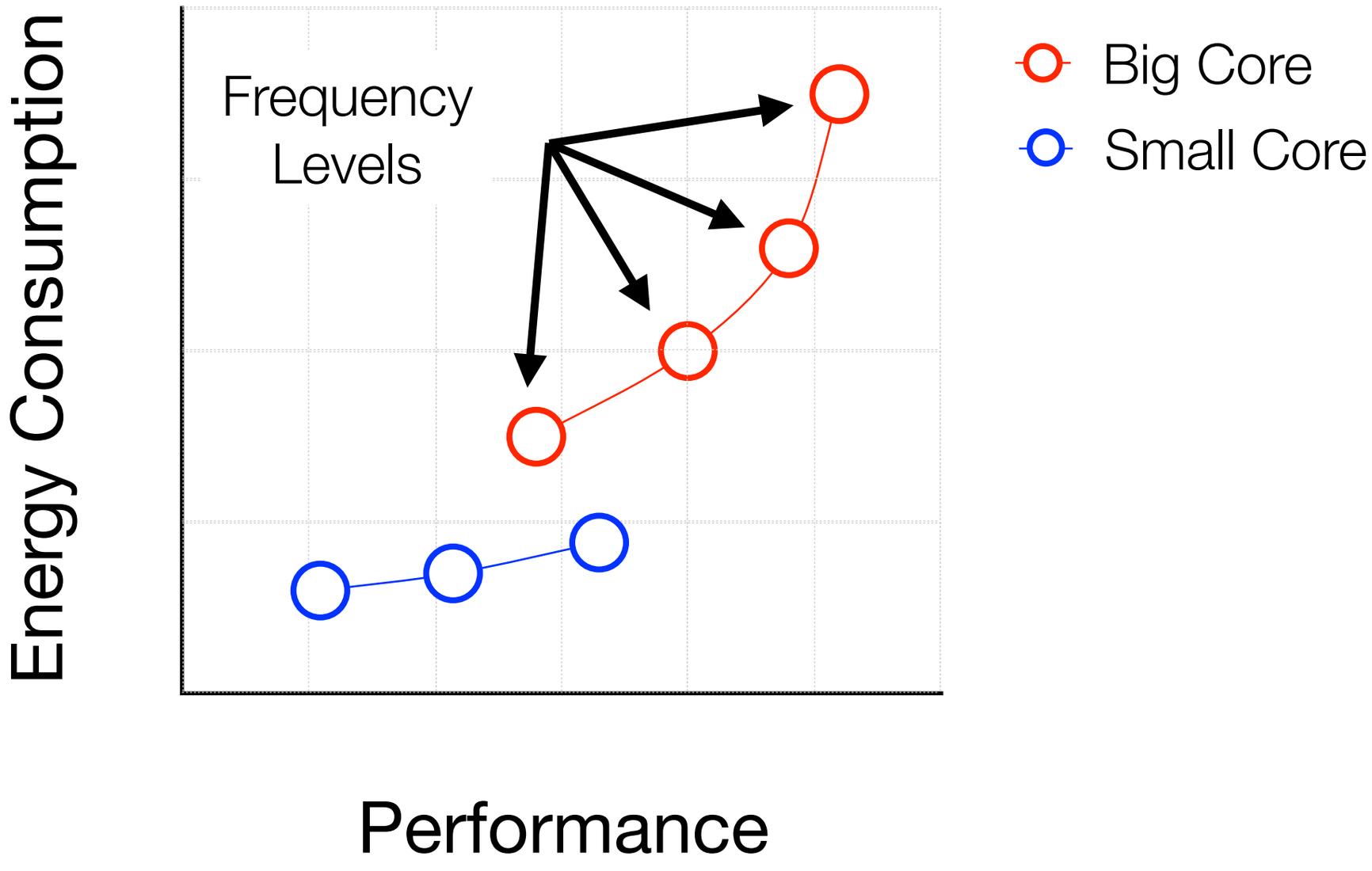
Asymmetric Chip-multiprocessor

- Offer a large performance-energy trade-off space
- Already used in commodity devices (e.g., Samsung Galaxy S6)



ACMP-based **GreenWeb** Runtime

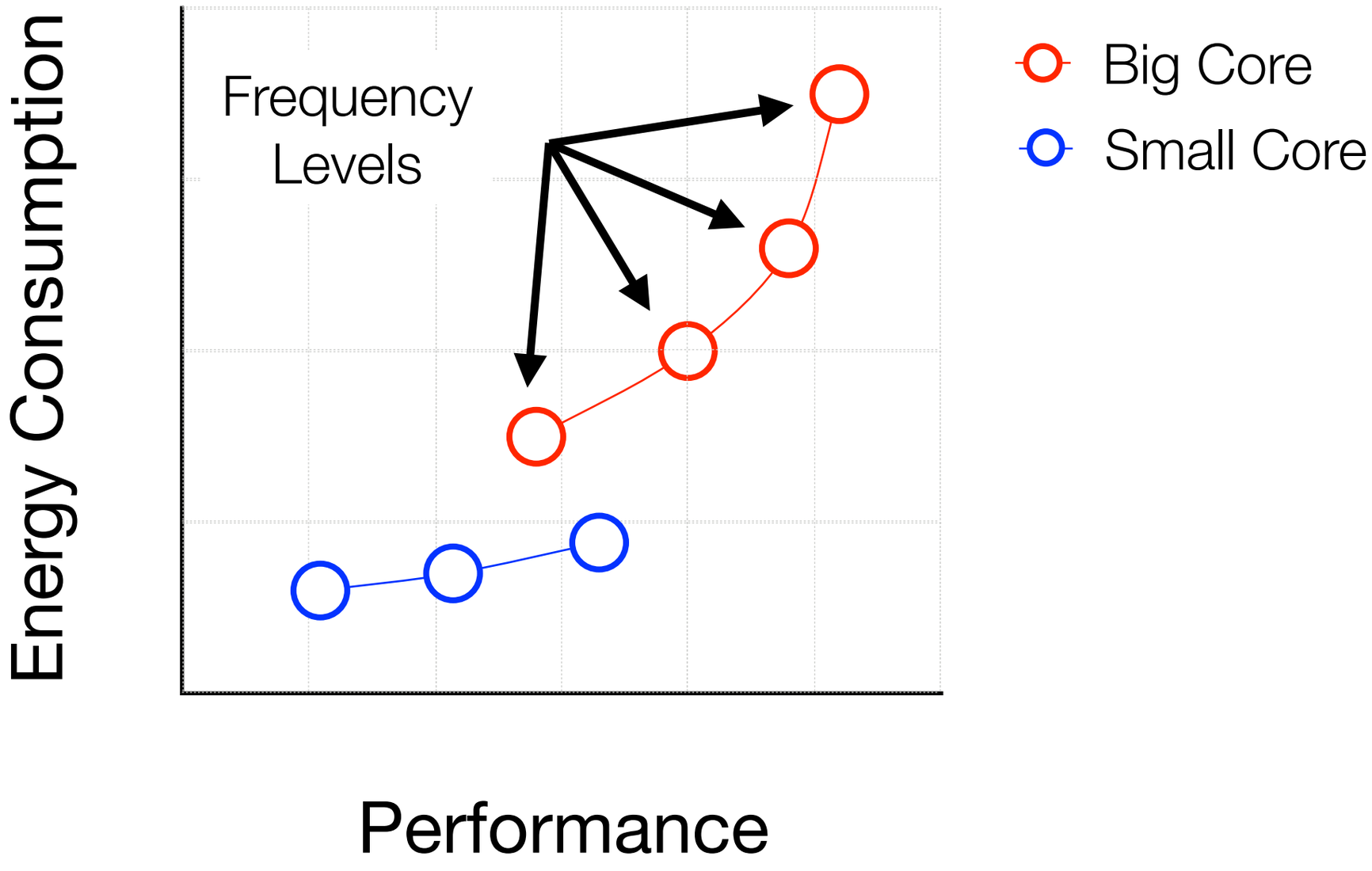
- Provide just enough energy to meet QoS constraints



ACMP-based **GreenWeb** Runtime

- Provide just enough energy to meet QoS constraints

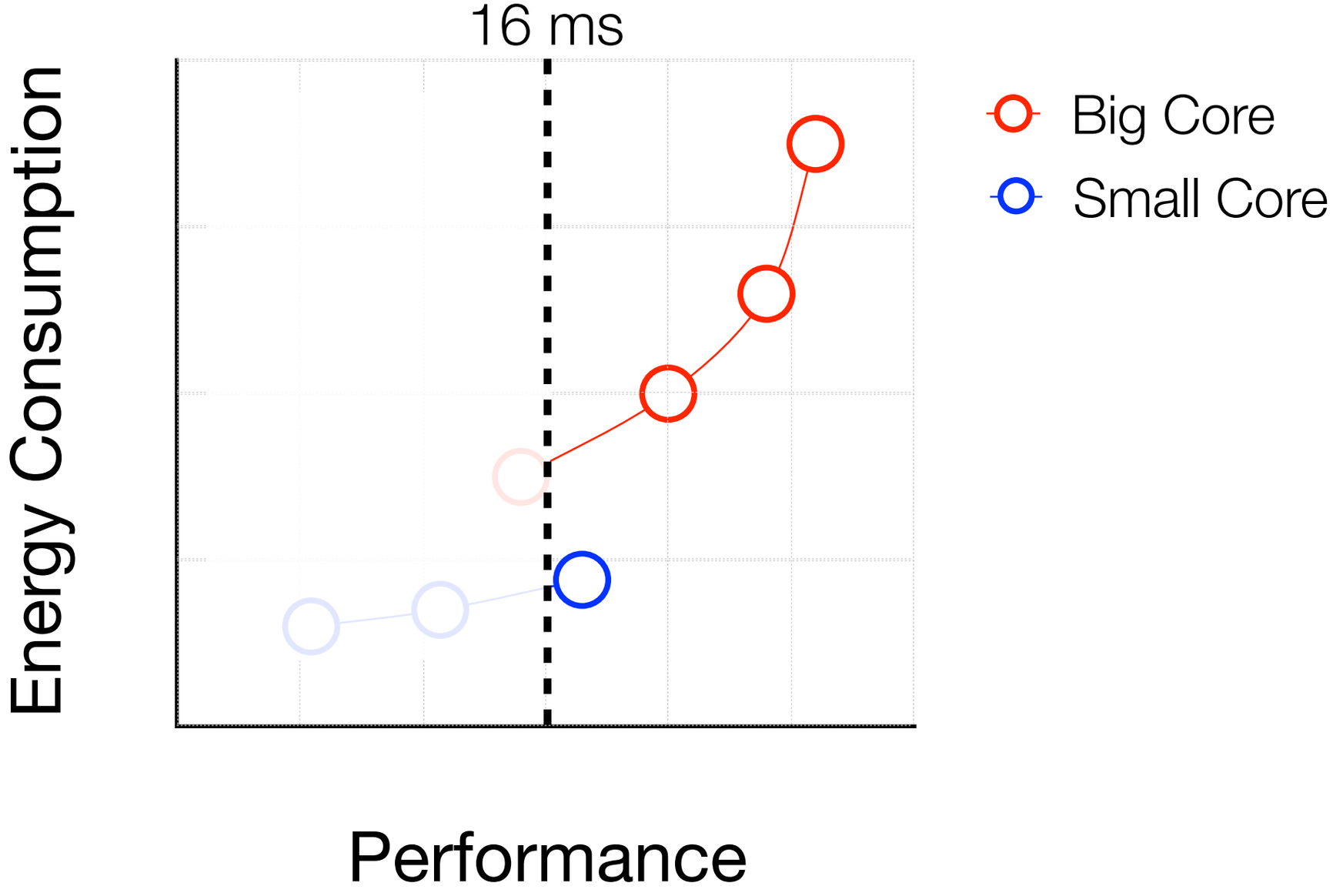
```
div {ontouchend: latency, 16 ms}
```



ACMP-based **GreenWeb** Runtime

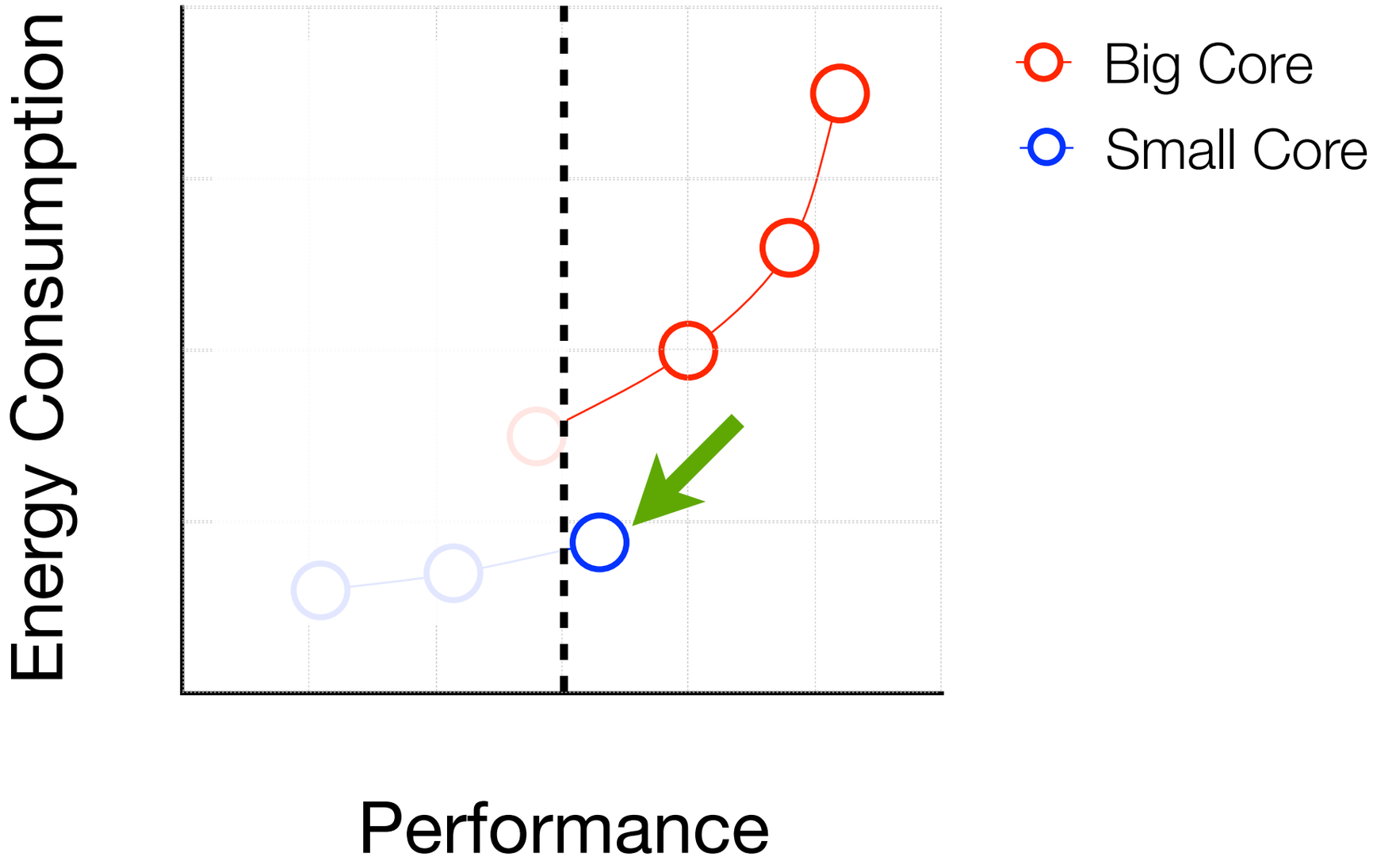
- Provide just enough energy to meet QoS constraints

```
div {ontouchend: latency, 16 ms}
```



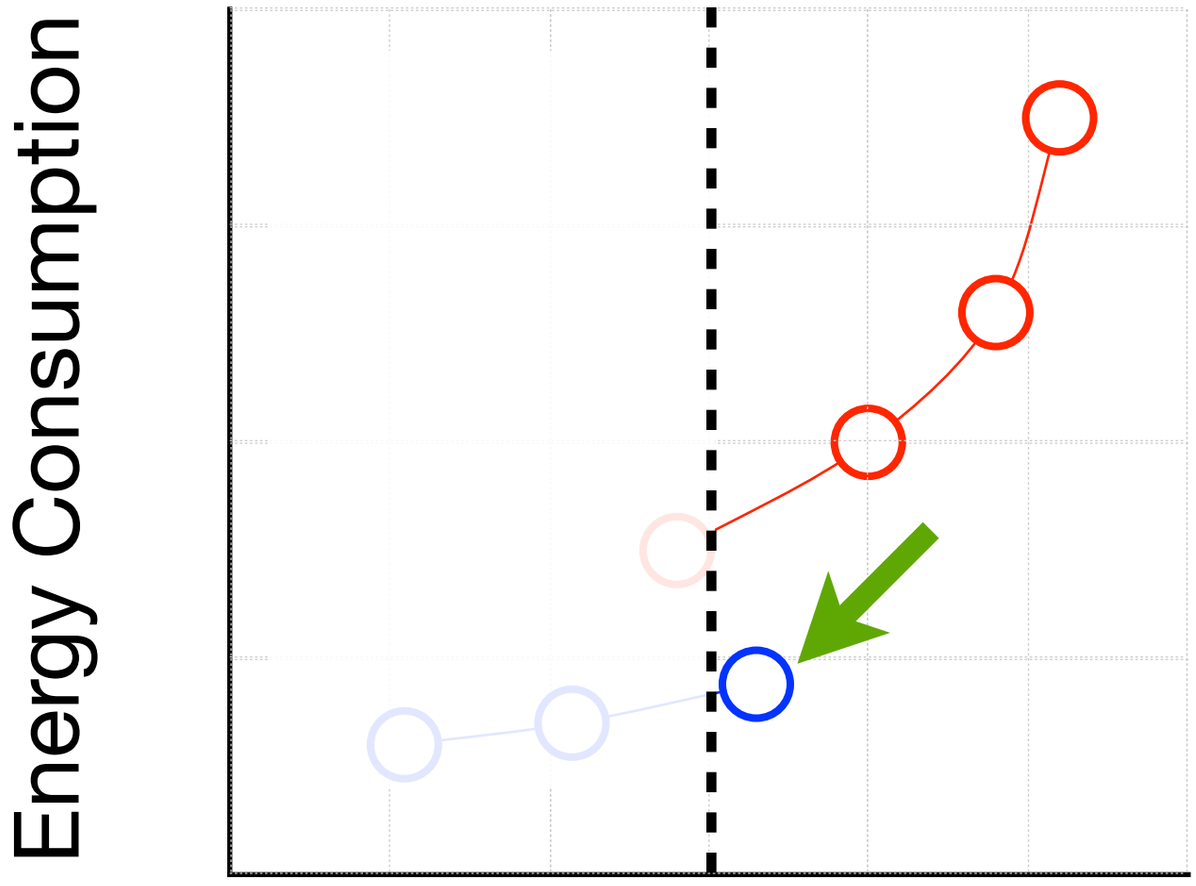
ACMP-based **GreenWeb** Runtime

- Provide just enough energy to meet QoS constraints



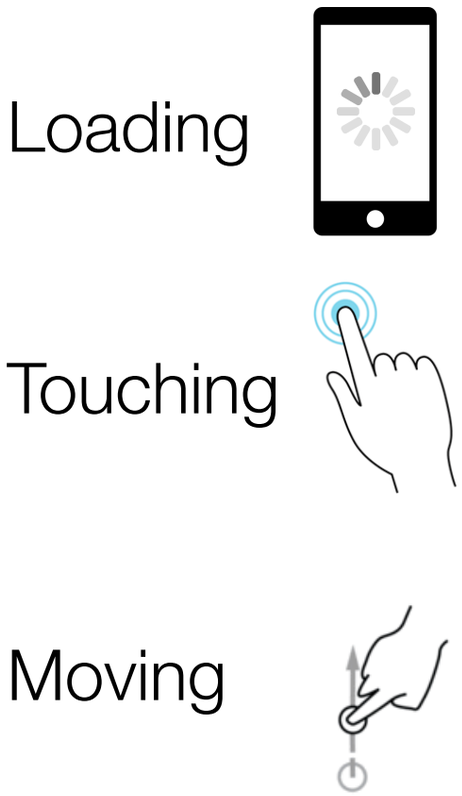
ACMP-based **GreenWeb** Runtime

- Provide just enough energy to meet QoS constraints

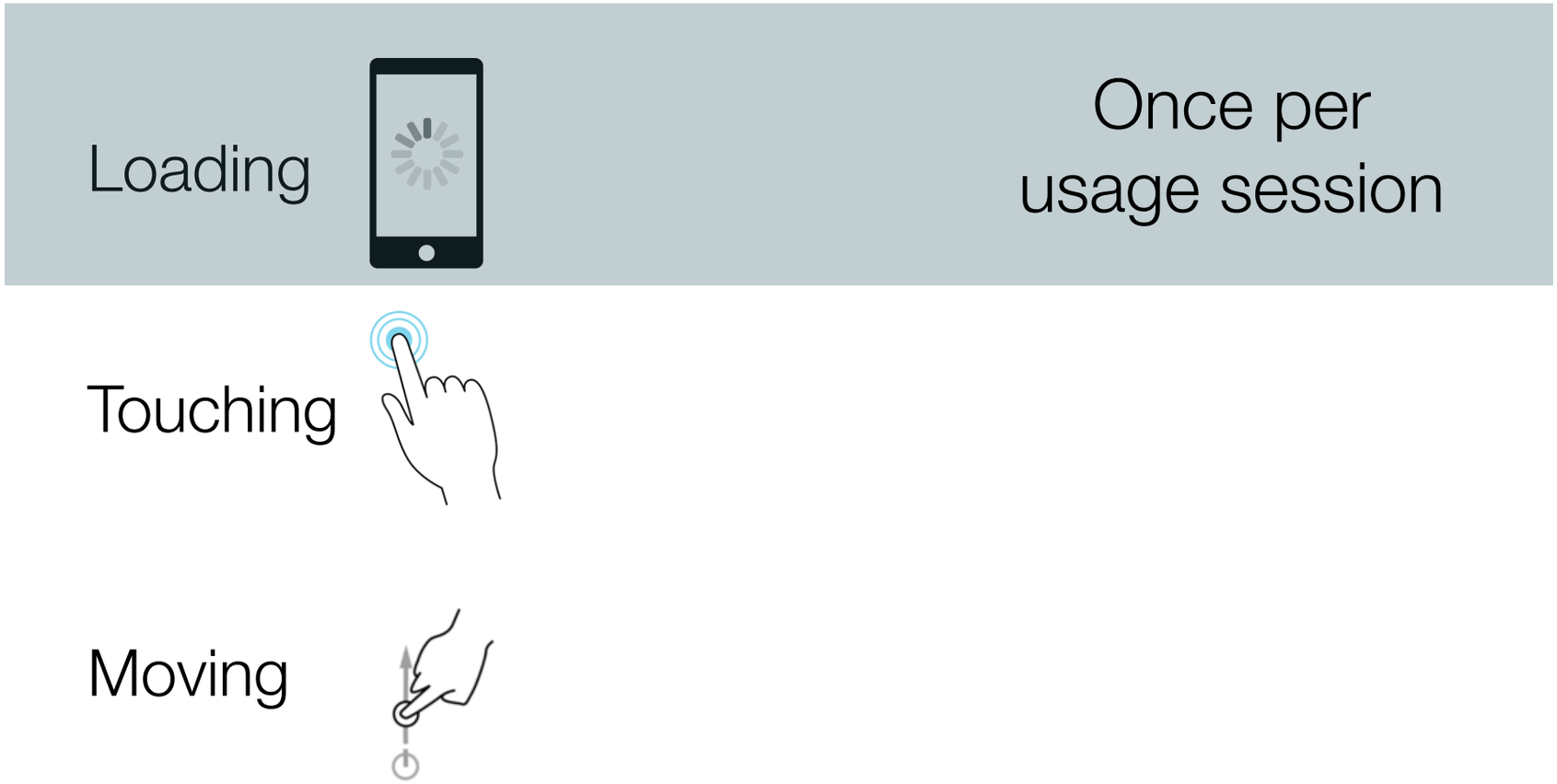


Predict the performance and energy of each configuration!

Different Strategies for Different Events



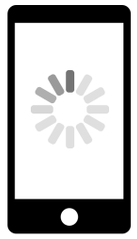
Different Strategies for Different Events



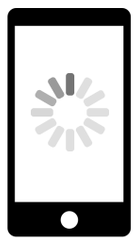
Different Strategies for Different Events

Events	WebRT Component
Loading 	Proactive Mechanism
Touching 	
Moving 	

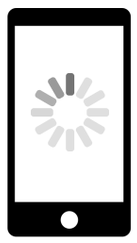
Different Strategies for Different Events

Events	WebRT Component
Loading 	Proactive Mechanism
Touching 	Repetitive in a usage session
Moving 	

Different Strategies for Different Events

Events	WebRT Component
Loading 	Proactive Mechanism
Touching 	Adaptive Mechanism
Moving 	

Different Strategies for Different Events

Events	WebRT Component
Loading 	Proactive Mechanism
Touching	Adaptive Mechanism
Moving	

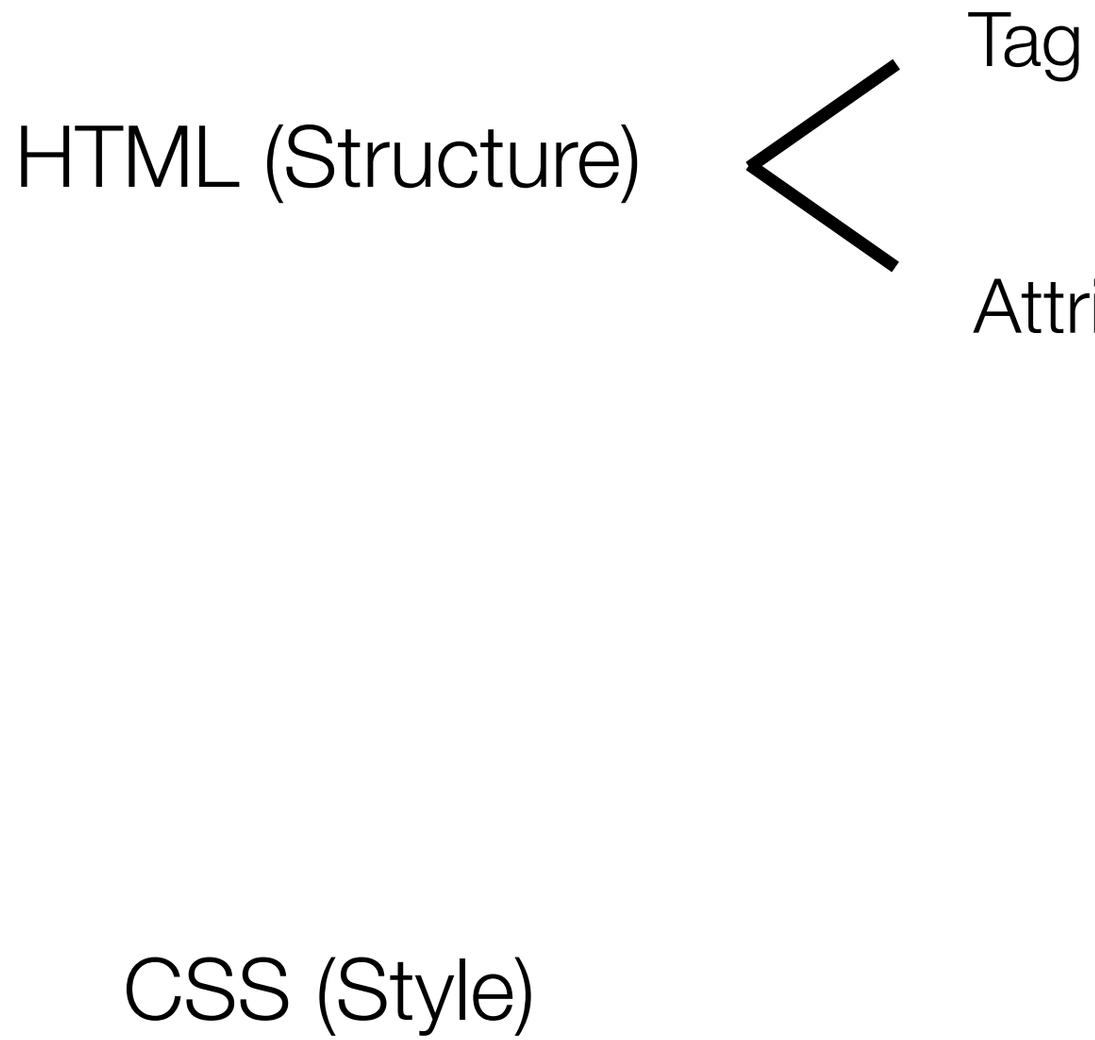
Breaking Down the Computations

Breaking Down the Computations

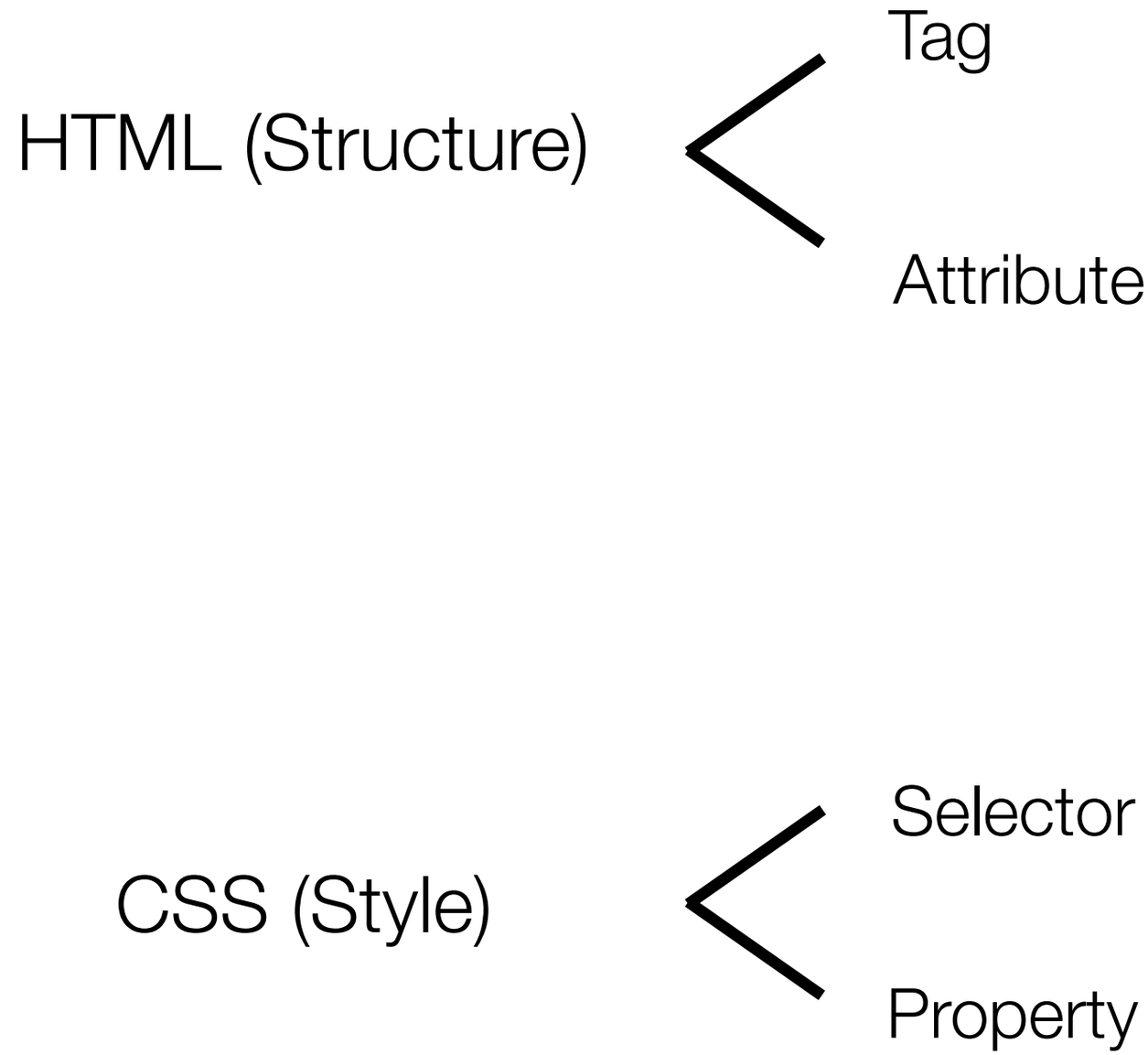
HTML (Structure)

CSS (Style)

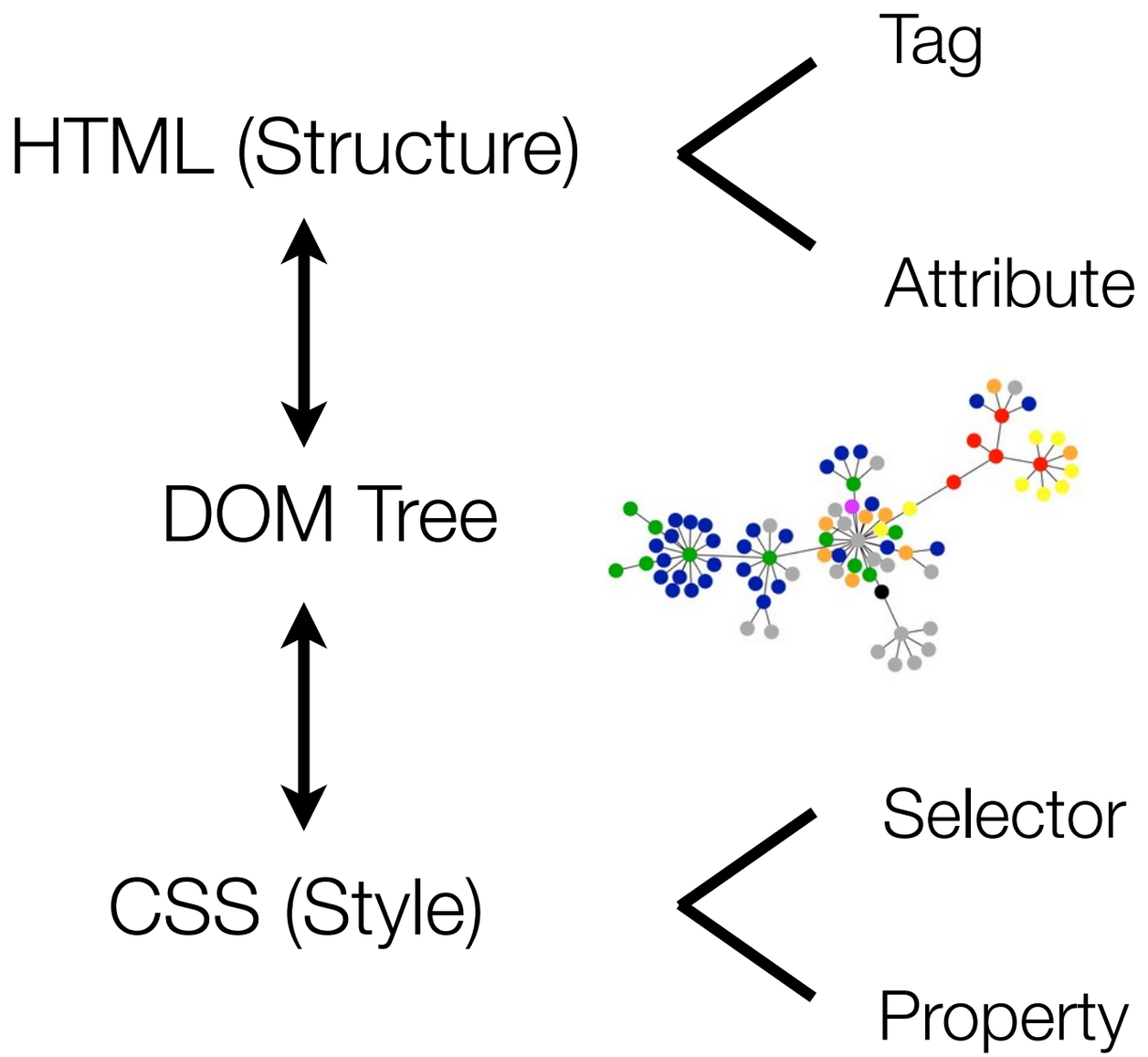
Breaking Down the Computations



Breaking Down the Computations



Breaking Down the Computations



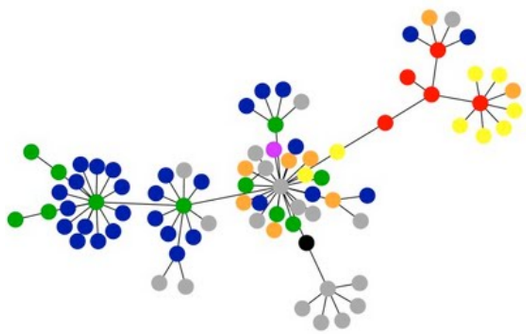
Breaking Down the Computations

Web Primitives

HTML (Structure)

Tag

Attribute



DOM Tree

CSS (Style)

Selector

Property

Predicting Loading Performance & Energy

Idea: predict load time & energy (responses)
based on Web primitives (predictors)

Predicting Loading Performance & Energy

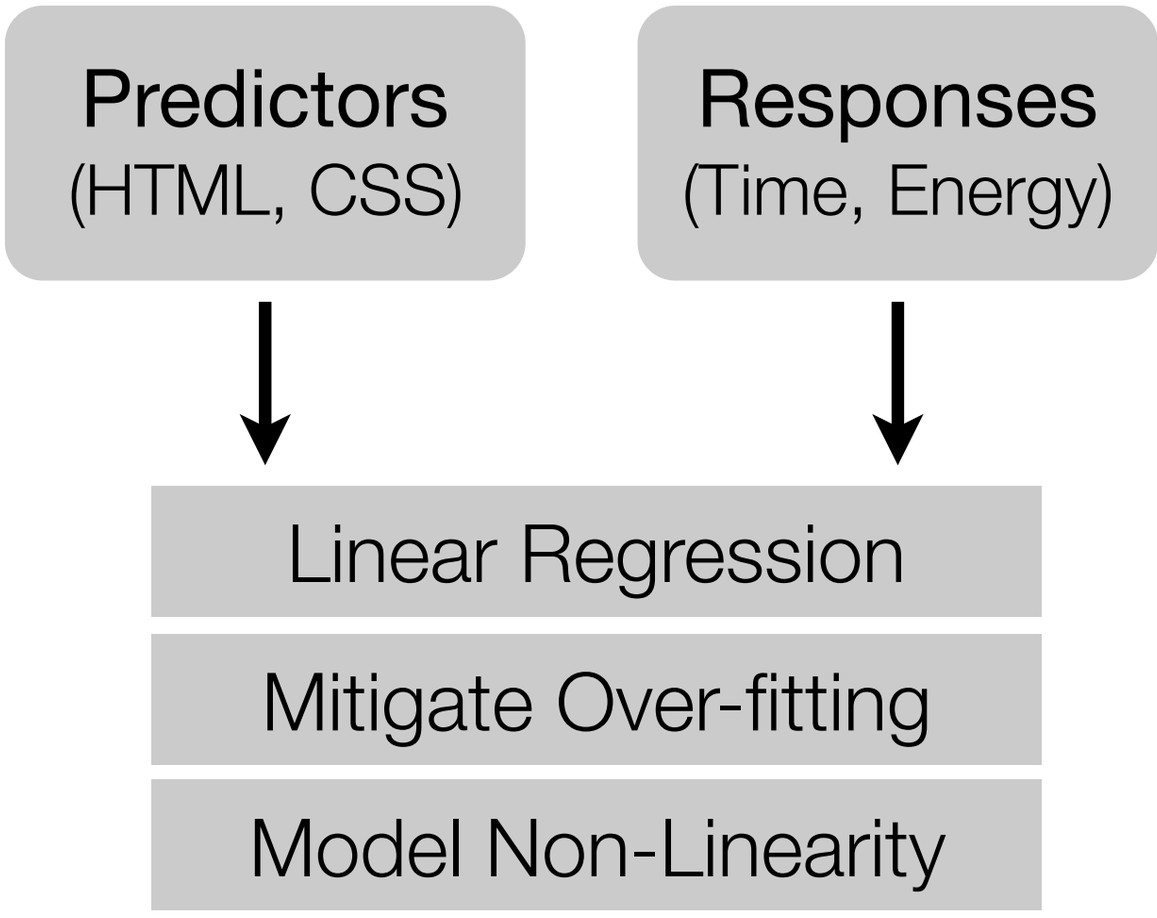
Identify Predictors
Training using top
2,500 webpages

Predictors
(HTML, CSS)

Responses
(Time, Energy)

Predicting Loading Performance & Energy

Identify Predictors
Training using top 2,500 webpages



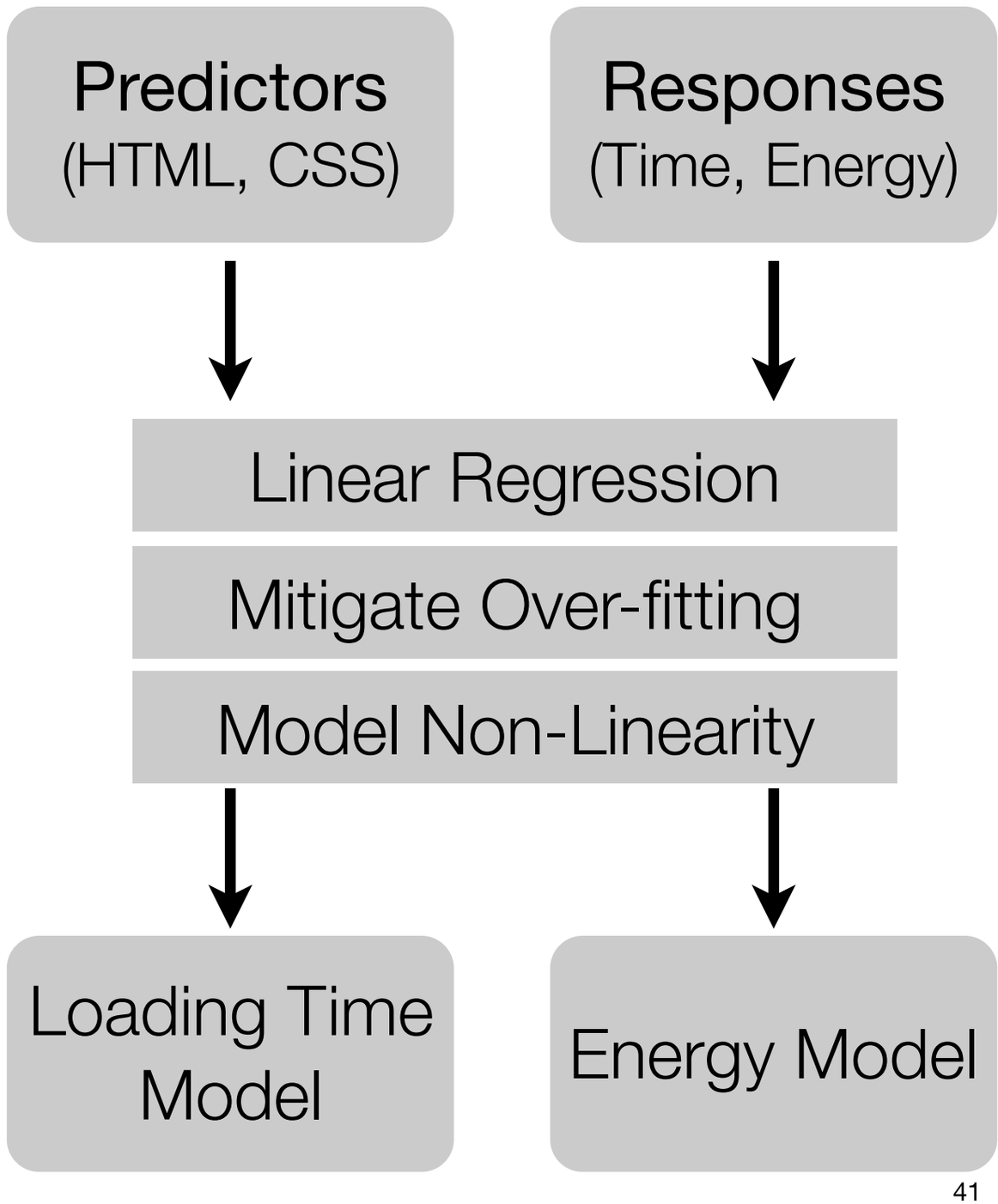
Model Construction & Refinement
Refine the linear model

Predicting Loading Performance & Energy

Identify Predictors
Training using top 2,500 webpages

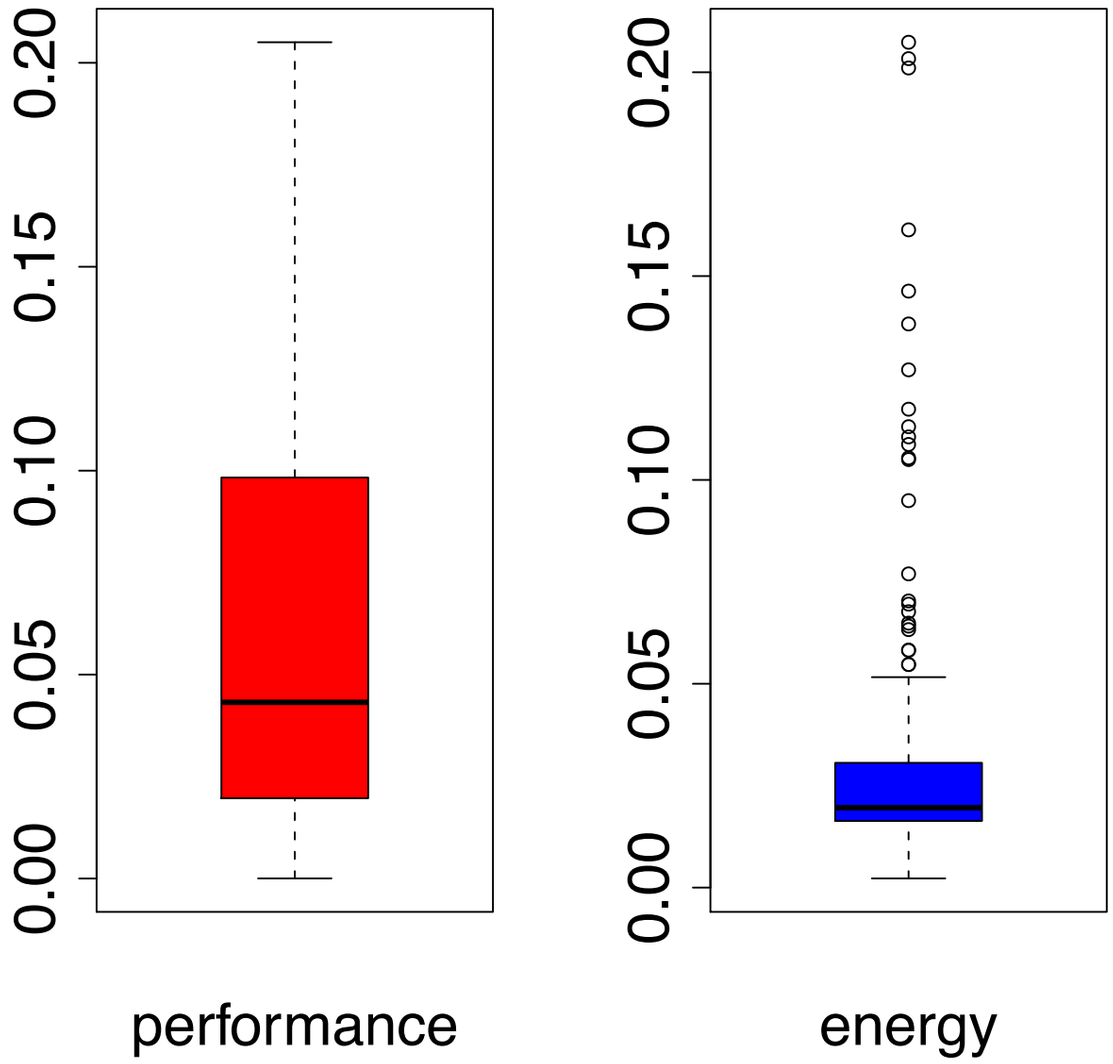
Model Construction & Refinement
Refine the linear model

Model Validation
Validating on another 2,500 webpages

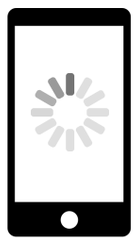


Predicting Loading Performance & Energy

Median prediction error is less than 5%



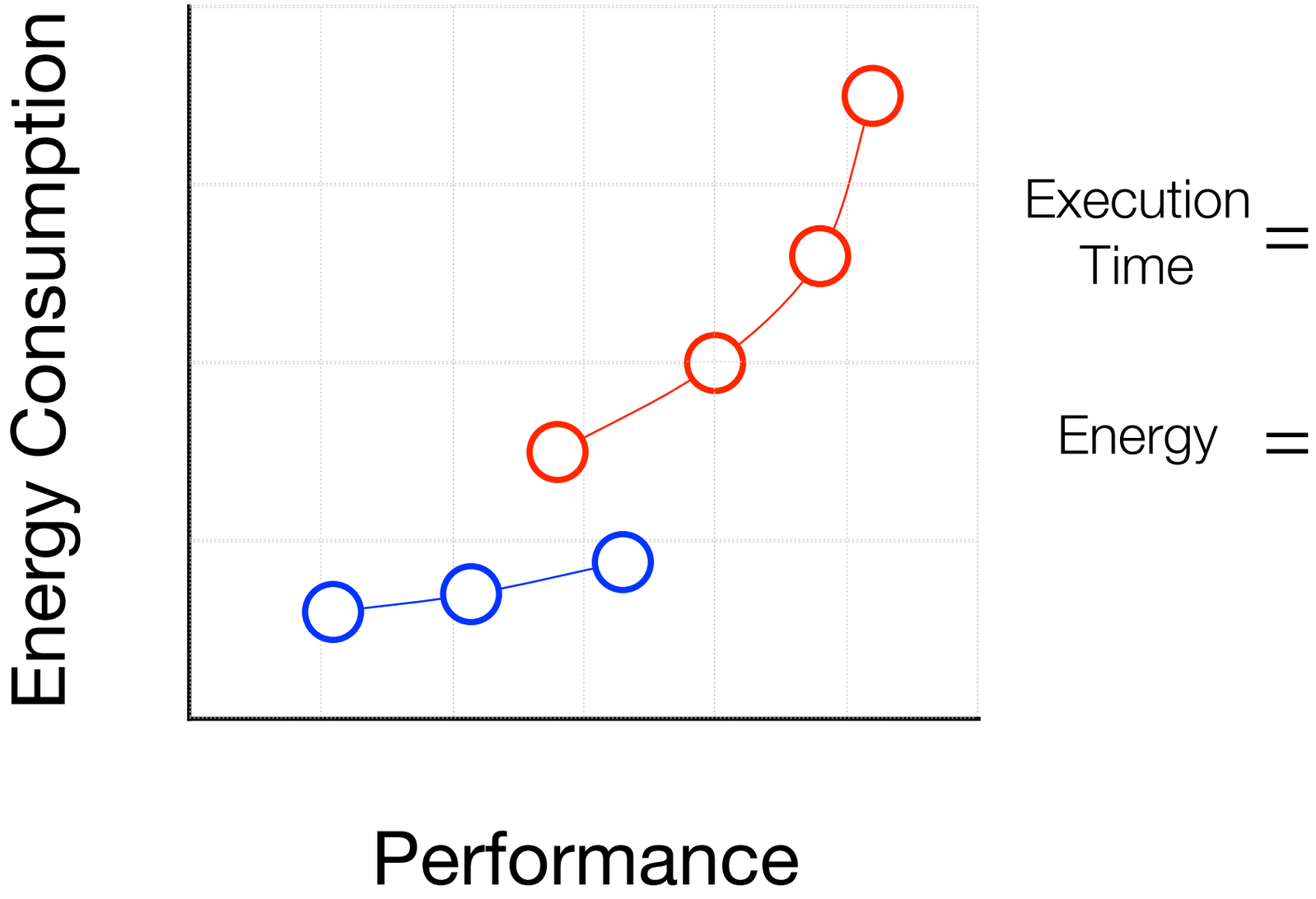
Different Strategies for Different Events

Interactions	WebRT Component
Loading 	Proactive Mechanism
Touching	Adaptive Mechanism
Moving	

Different Strategies for Different Events

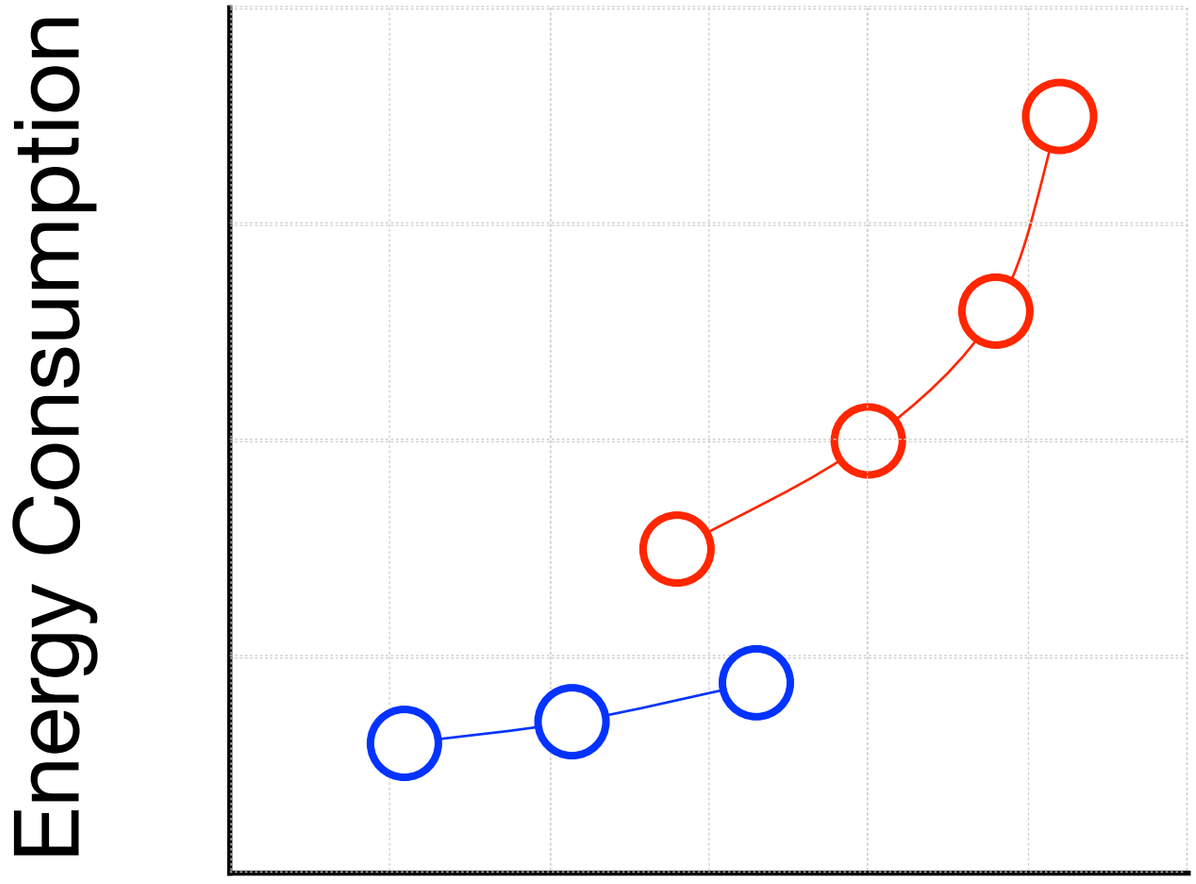
Interactions	WebRT Component
Loading	Proactive Mechanism
Touching 	Adaptive Mechanism
Moving 	

ACMP-based **GreenWeb** Runtime



[PLDI 2003] Xie, et al., “Compile-Time Dynamic Voltage Scaling Settings: Opportunities and Limits”

ACMP-based **GreenWeb** Runtime

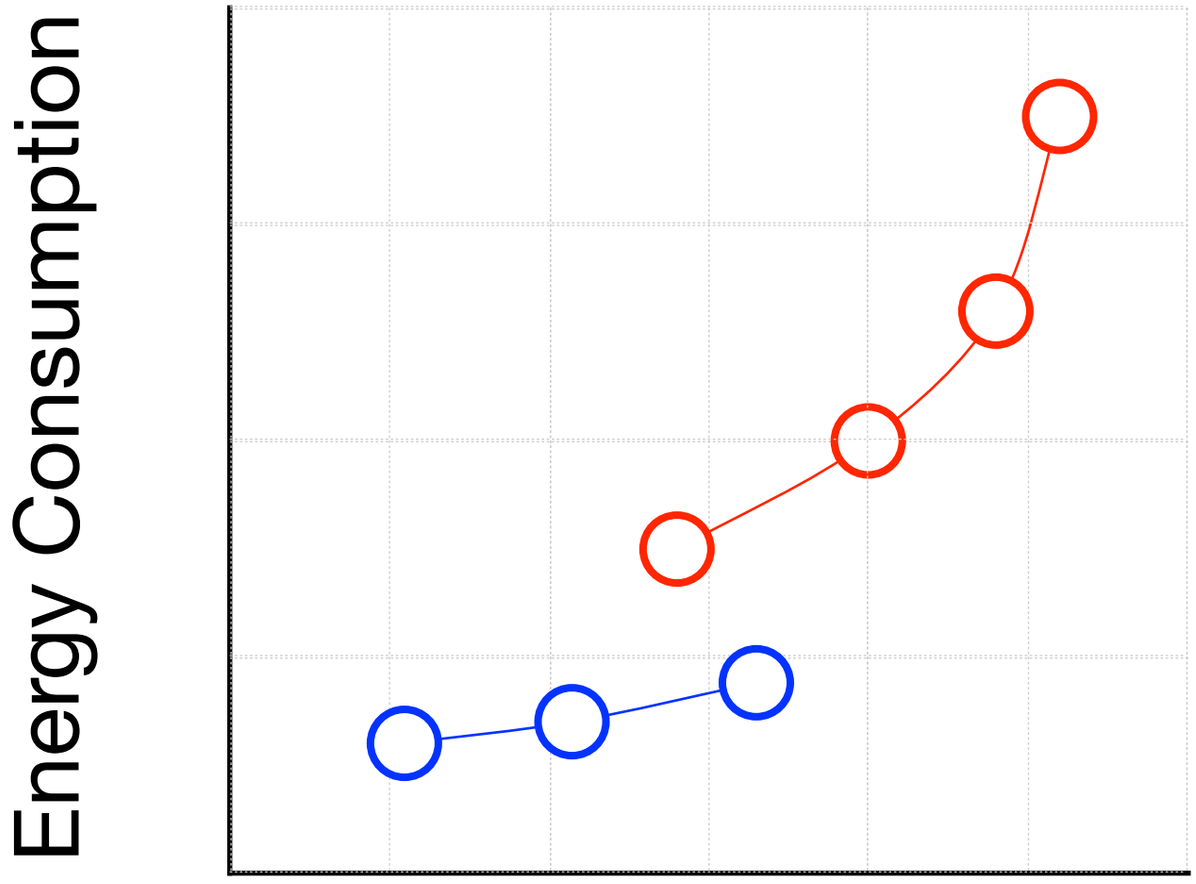


Execution Time = T_{memory} +

Energy =

[PLDI 2003] Xie, et al., “Compile-Time Dynamic Voltage Scaling Settings: Opportunities and Limits”

ACMP-based **GreenWeb** Runtime

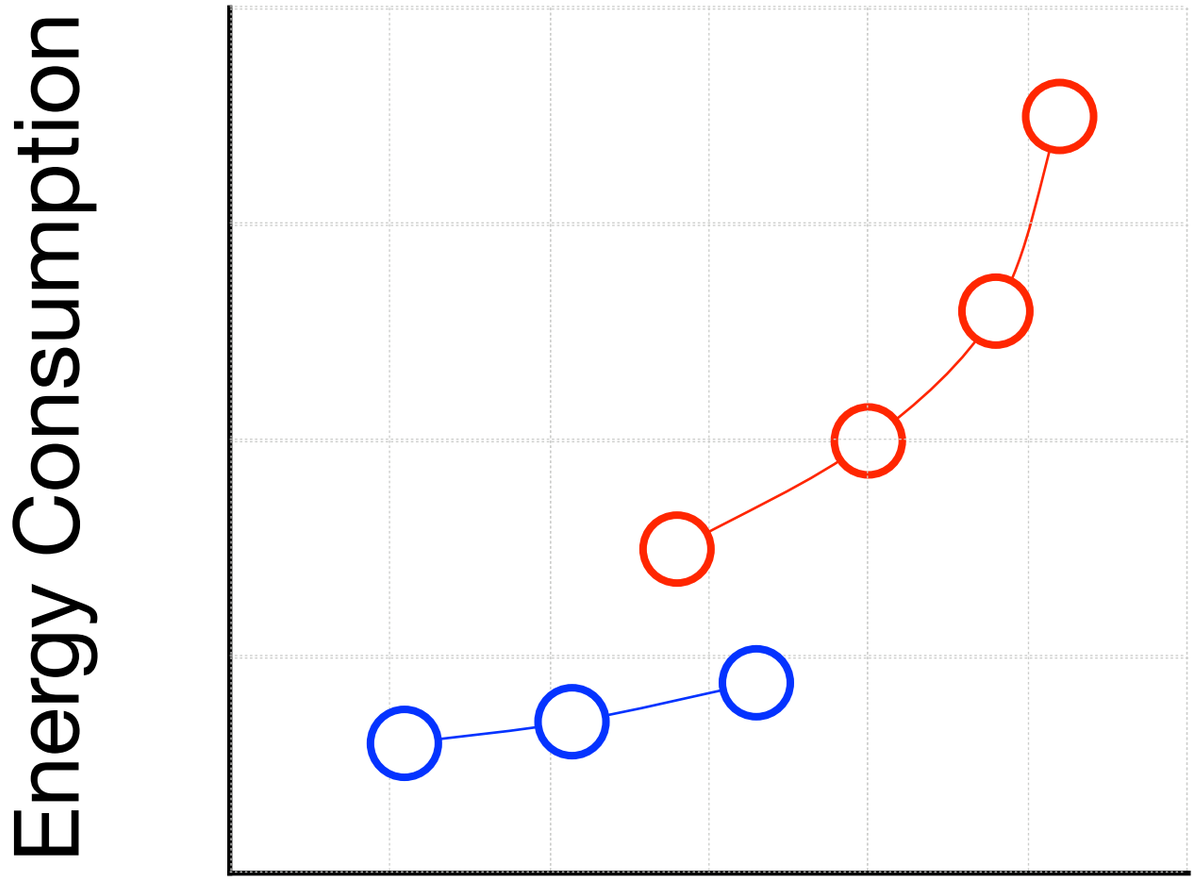


$$\text{Execution Time} = T_{memory} + T_{cpu}$$

$$\text{Energy} =$$

[PLDI 2003] Xie, et al., “Compile-Time Dynamic Voltage Scaling Settings: Opportunities and Limits”

ACMP-based **GreenWeb** Runtime

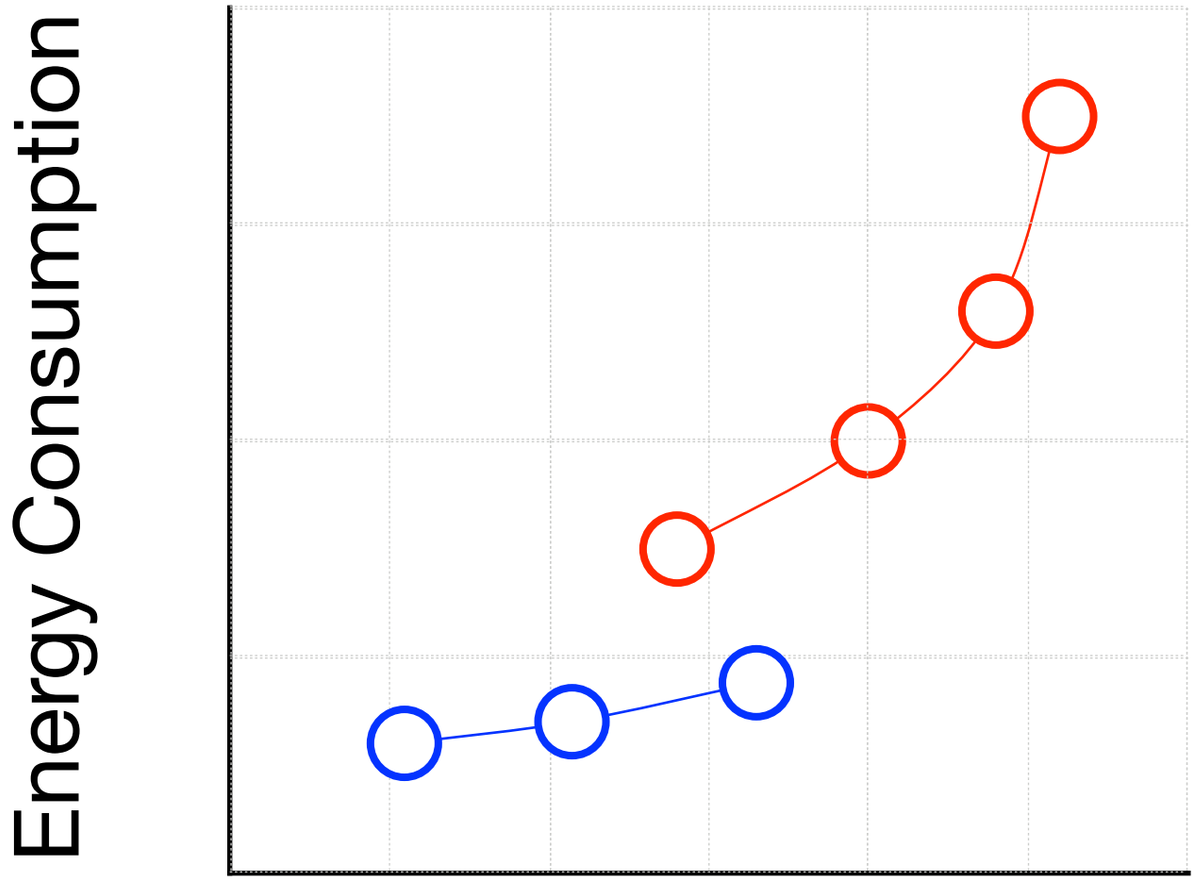


$$\text{Execution Time} = T_{memory} + N_{cycles} / f$$

$$\text{Energy} =$$

[PLDI 2003] Xie, et al., “Compile-Time Dynamic Voltage Scaling Settings: Opportunities and Limits”

ACMP-based **GreenWeb** Runtime

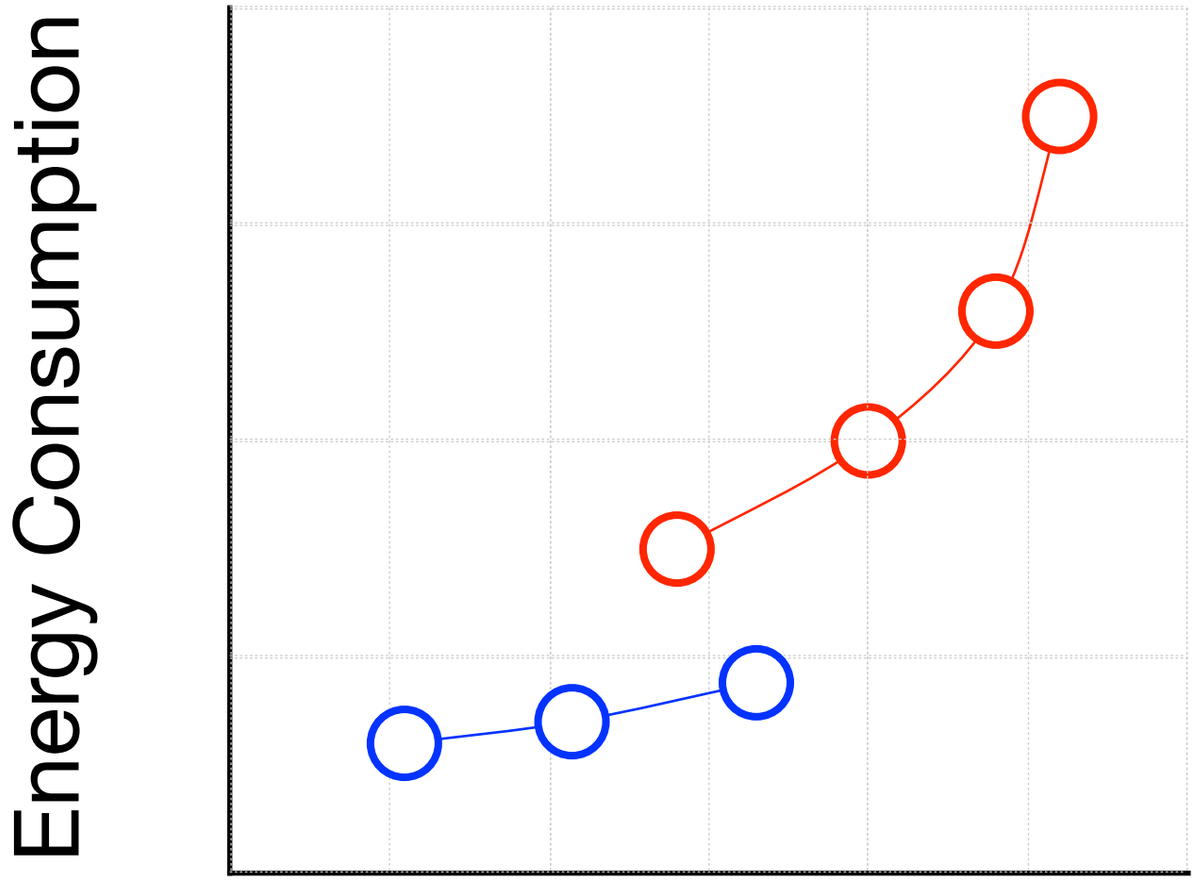


$$\text{Execution Time} = T_{memory} + N_{cycles} / f$$

$$\text{Energy} =$$

[PLDI 2003] Xie, et al., “Compile-Time Dynamic Voltage Scaling Settings: Opportunities and Limits”

ACMP-based **GreenWeb** Runtime

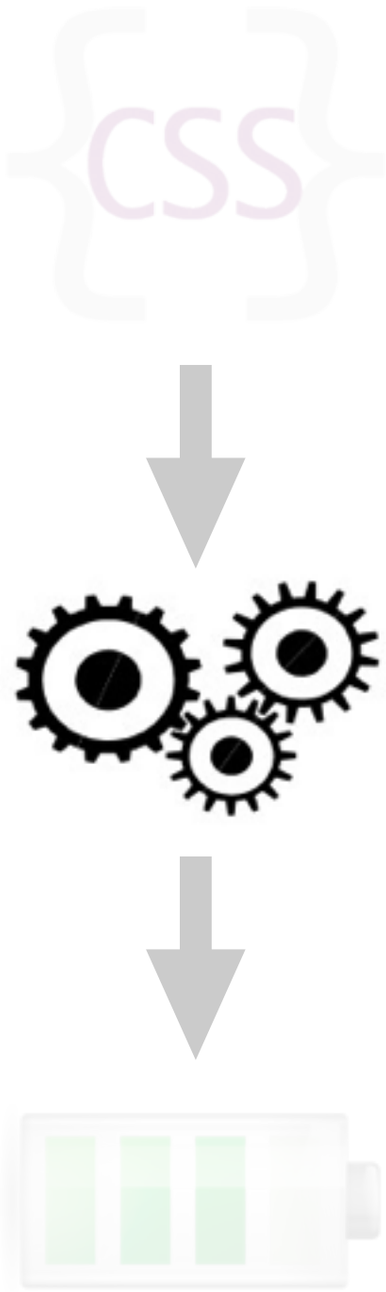


$$\text{Execution Time} = T_{memory} + N_{cycles} / f$$

$$\text{Energy} = \text{Execution Time} \times \text{Power}$$

[PLDI 2003] Xie, et al., “Compile-Time Dynamic Voltage Scaling Settings: Opportunities and Limits”

GreenWeb: Language for Energy-Efficiency



- Language abstractions
- **Runtime** that saves energy while meeting the QoS constraints
- Result
  hardware/software implementations

GreenWeb: Language for Energy-Efficiency

- Language abstractions

- Runtime
  the QoS constraints

- **Result** in 60% energy savings on real hardware/software implementations

Real Hardware/Software Setup

ODroid XU+E development board, which contains an Exynos 5410 SoC used in Samsung Galaxy S4.

Real Hardware/Software Setup

ODroid XU+E development board, which contains an Exynos 5410 SoC used in Samsung Galaxy S4.

Big core cluster: ARM Cortex A15, OoO with 3 issue

Little core cluster: ARM Cortex A7, In-order with 2 issue

Overhead:

- Frequency switch: 100 us
- Core migration: 20 us

Real Hardware/Software Setup

ODroid XU+E development board, which contains an Exynos 5410 SoC used in Samsung Galaxy S4.

Big core cluster: ARM Cortex A15, OoO with 3 issue

Little core cluster: ARM Cortex A7, In-order with 2 issue

Overhead:

- Frequency switch: 100 us
- Core migration: 20 us

Implementation incorporated into Chrome running on Android.

Real Hardware/Software Setup

ODroid XU+E development board, which contains an Exynos 5410 SoC used in Samsung Galaxy S4.

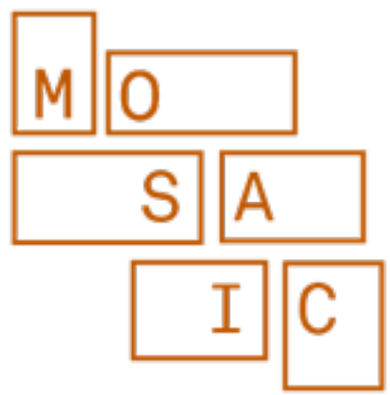
Big core cluster: ARM Cortex A15, OoO with 3 issue

Little core cluster: ARM Cortex A7, In-order with 2 issue

Overhead:

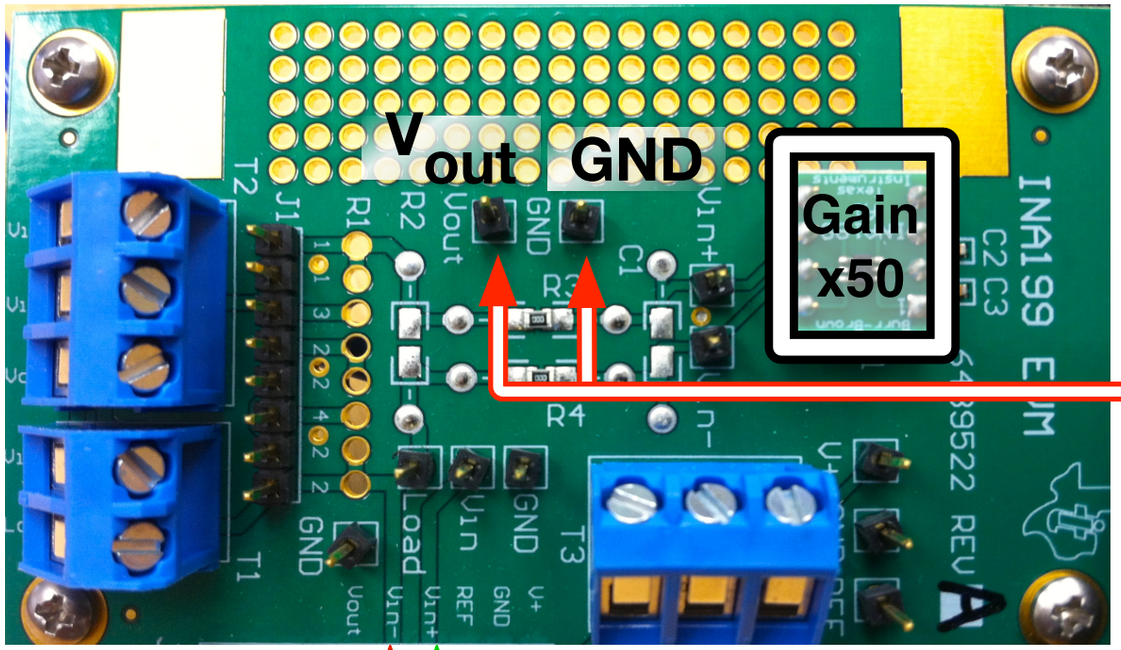
- Frequency switch: 100 us
- Core migration: 20 us

Implementation incorporated into Chrome running on Android.



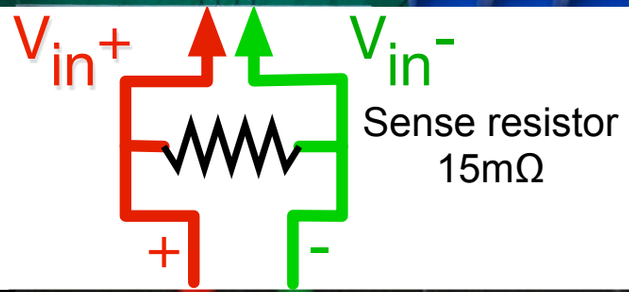
UI-level record and replay for reproducibility. [ISPASS'15]

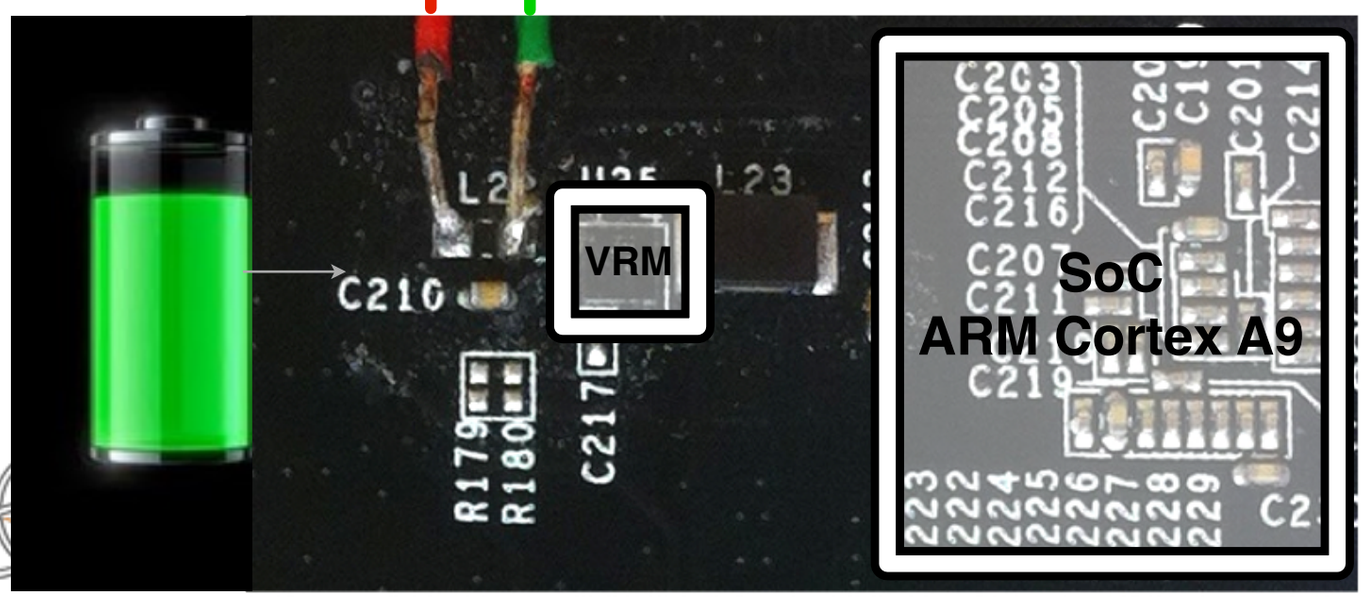
Power and Energy Measurements



Probe

Data Acquisition
(DAQ)





$$\text{Power} = (V_{in}^{+} - V_{in}^{-}) / R_{sense} * V_{in}^{-}$$

Evaluation

- Baseline Mechanisms
  - Highest performance (**Perf**) — Standard to guarantee responsiveness
  - Interactive governor (**Interactive**) — Android default

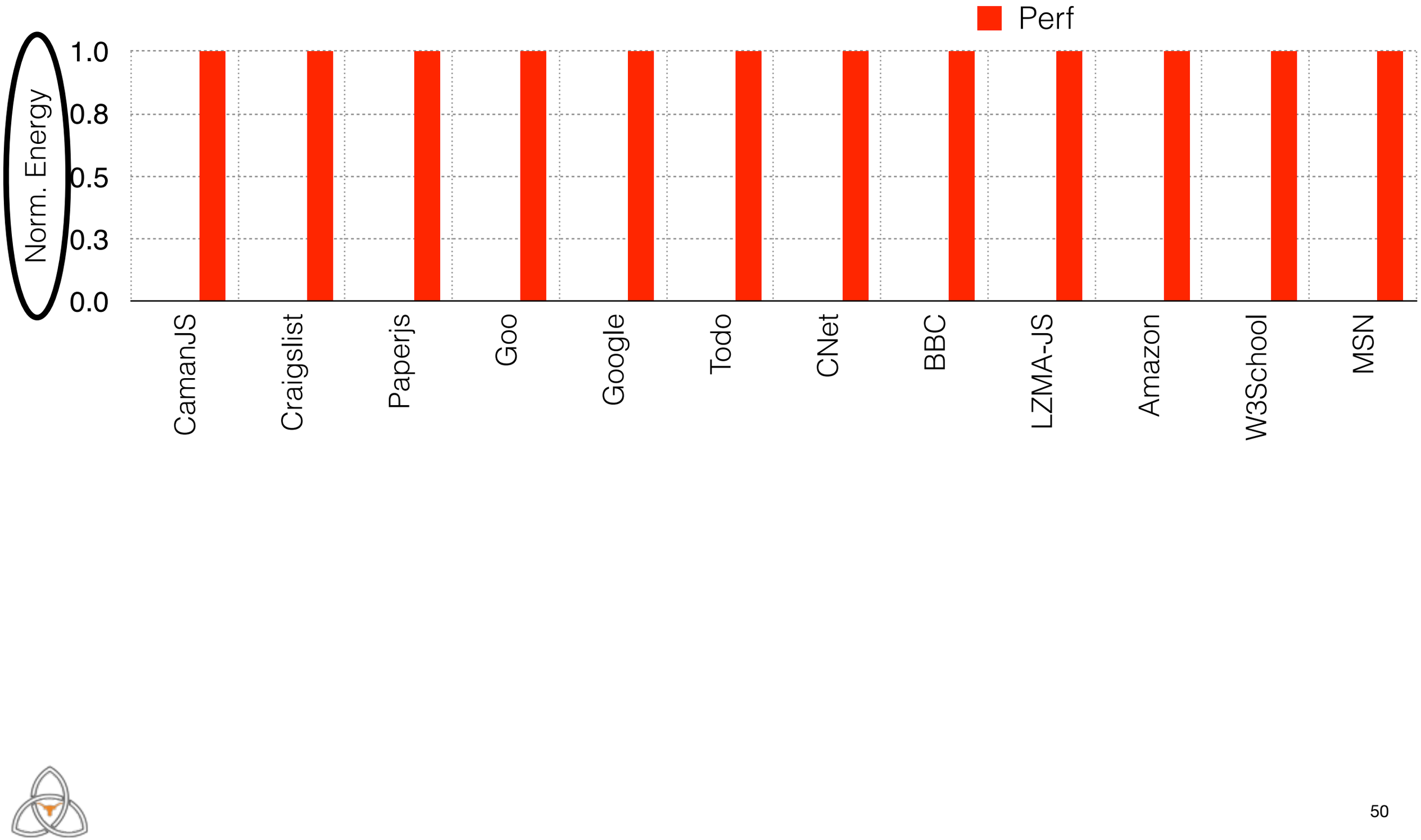
Evaluation

- Baseline Mechanisms
  - Highest performance (**Perf**) — Standard to guarantee responsiveness
  - Interactive governor (**Interactive**) — Android default
- Metrics
  - Energy Saving
  - QoS Violation

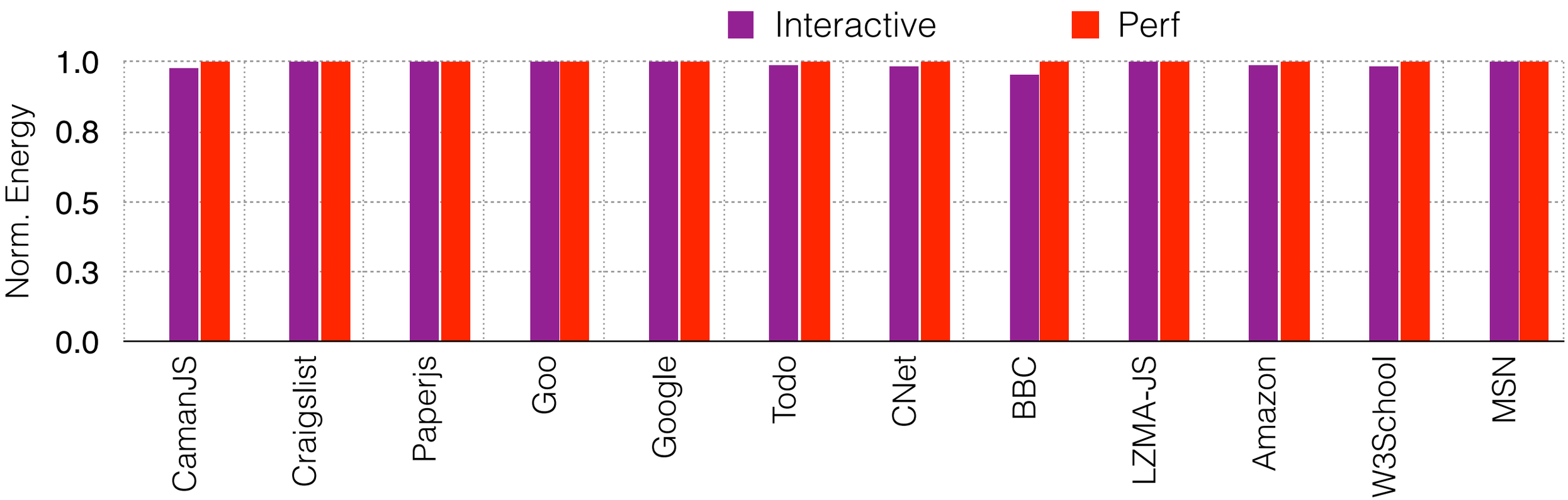
Evaluation

- Baseline Mechanisms
  - Highest performance (**Perf**) — Standard to guarantee responsiveness
  - Interactive governor (**Interactive**) — Android default
- Metrics
  - Energy Saving
  - QoS Violation
- Applications
  - Top webpages (e.g., www.amazon.com)
  - Web Apps based on popular frameworks (e.g., Todo List)

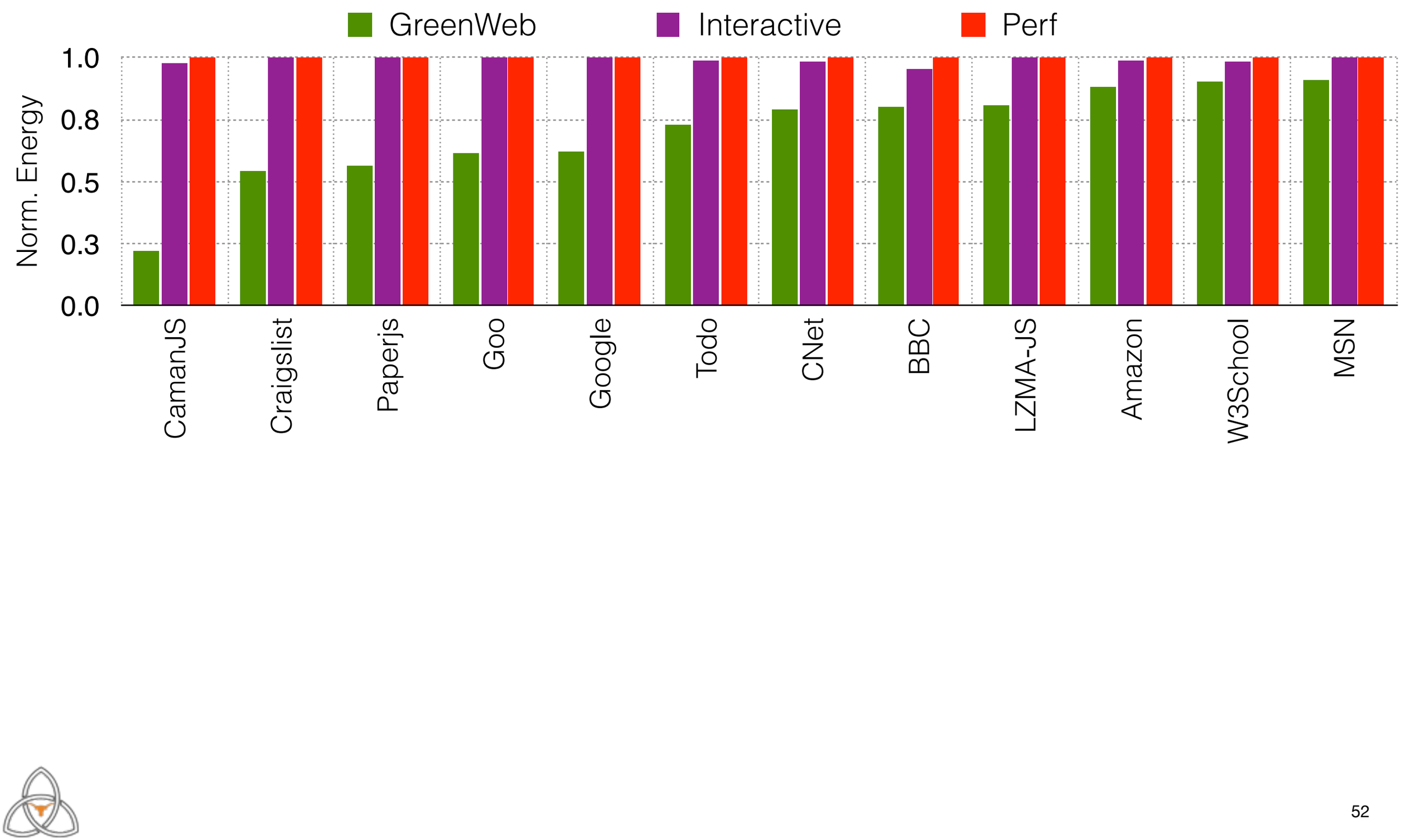
Evaluation Results



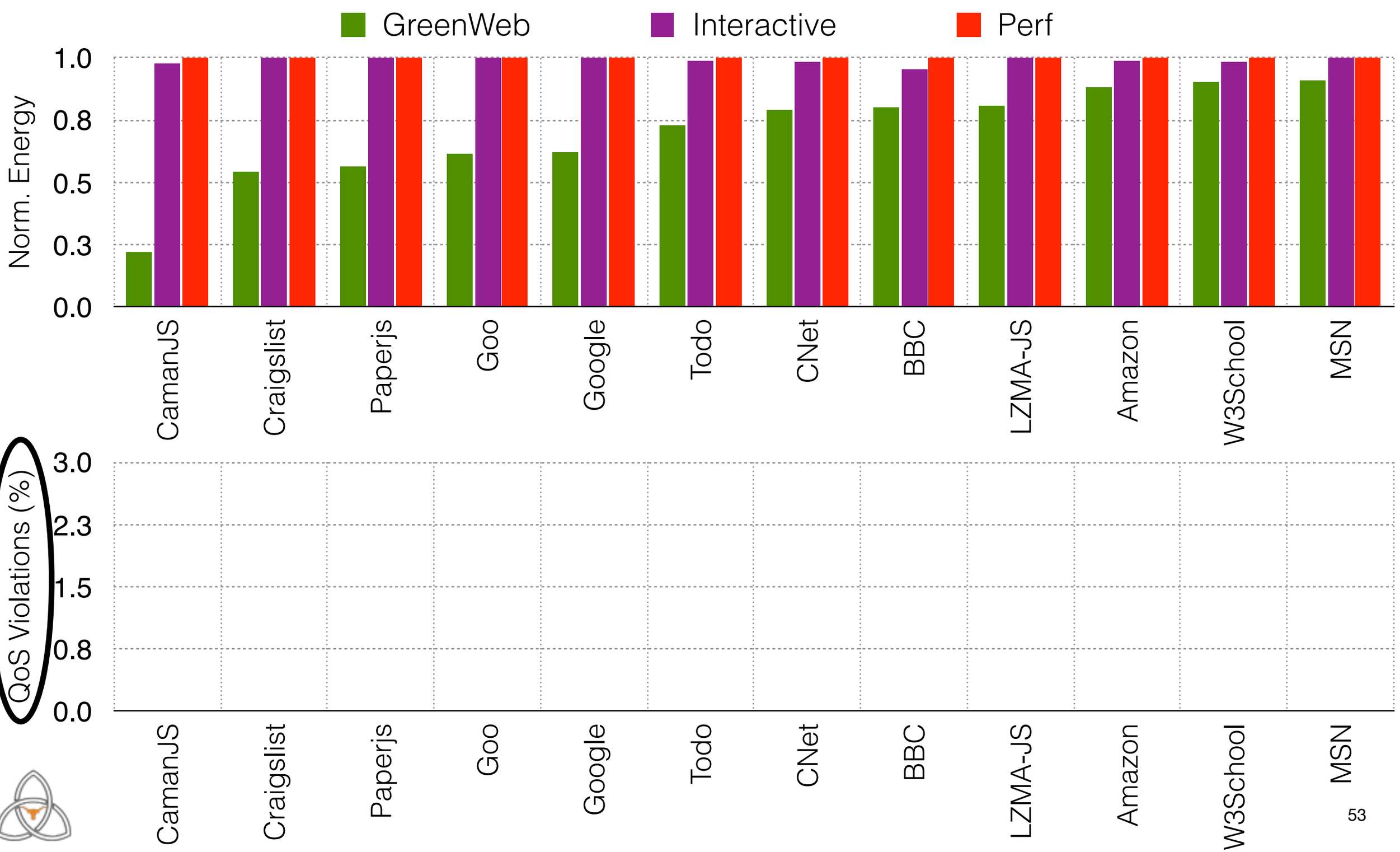
Evaluation Results



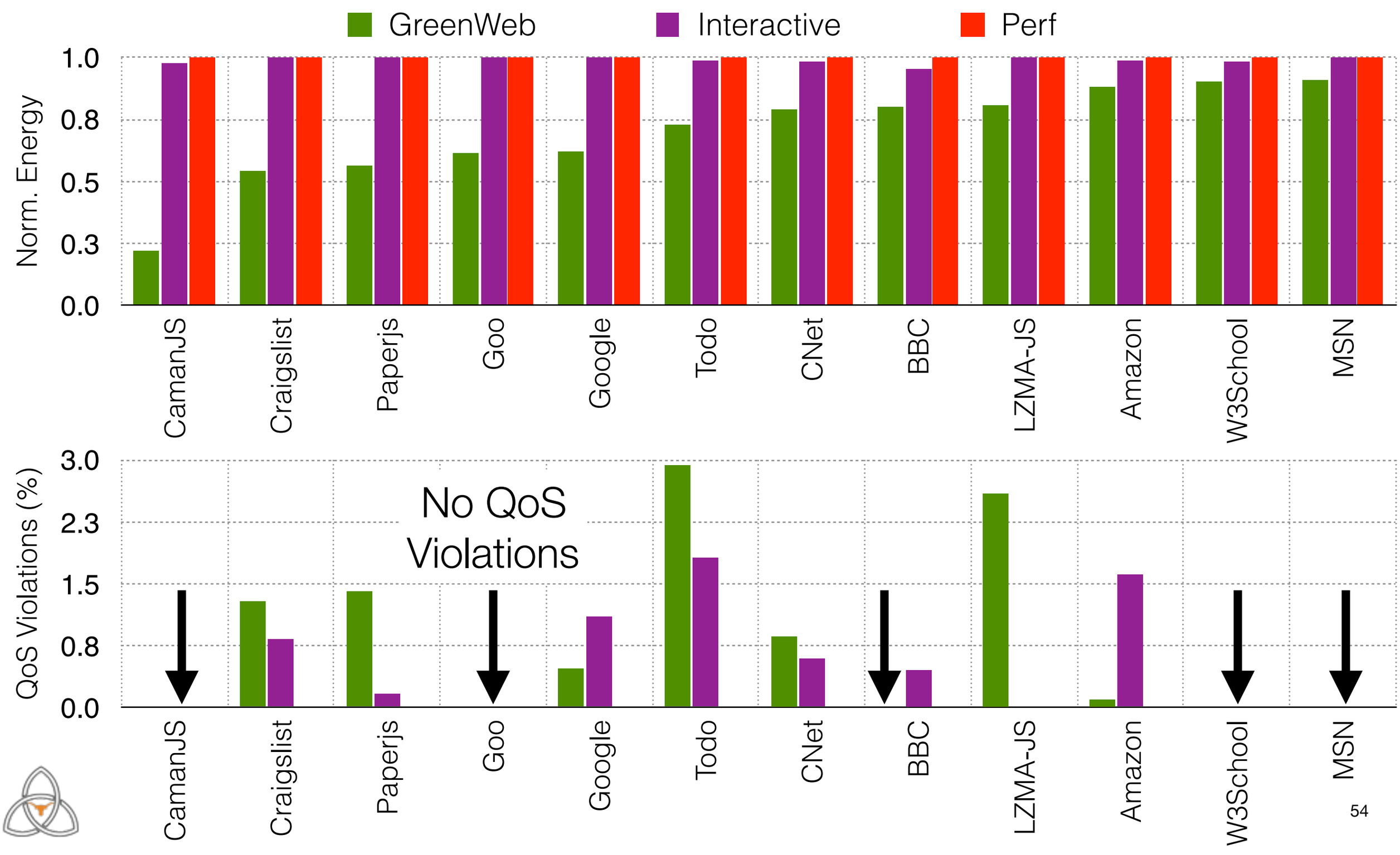
Evaluation Results



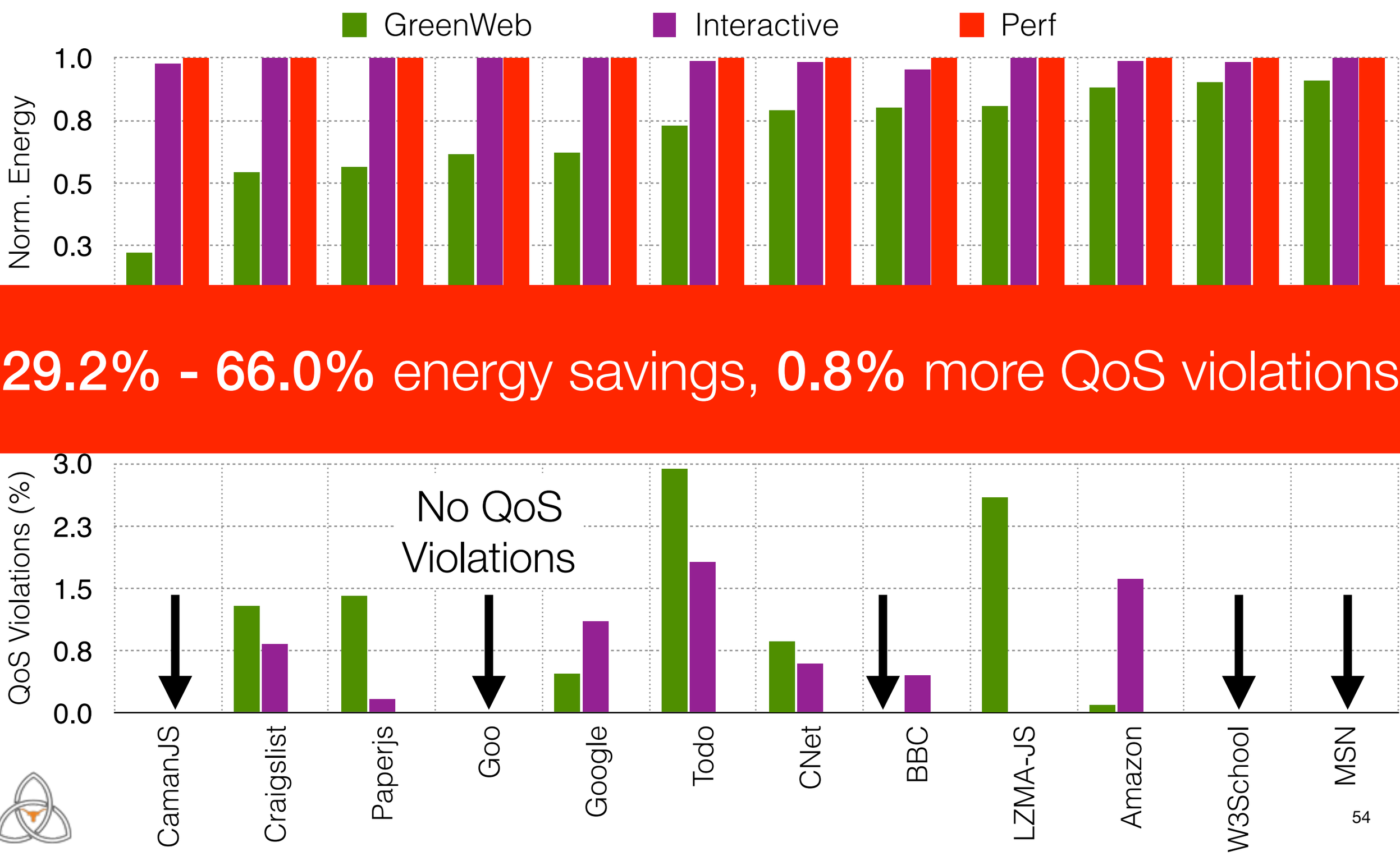
Evaluation Results



Evaluation Results

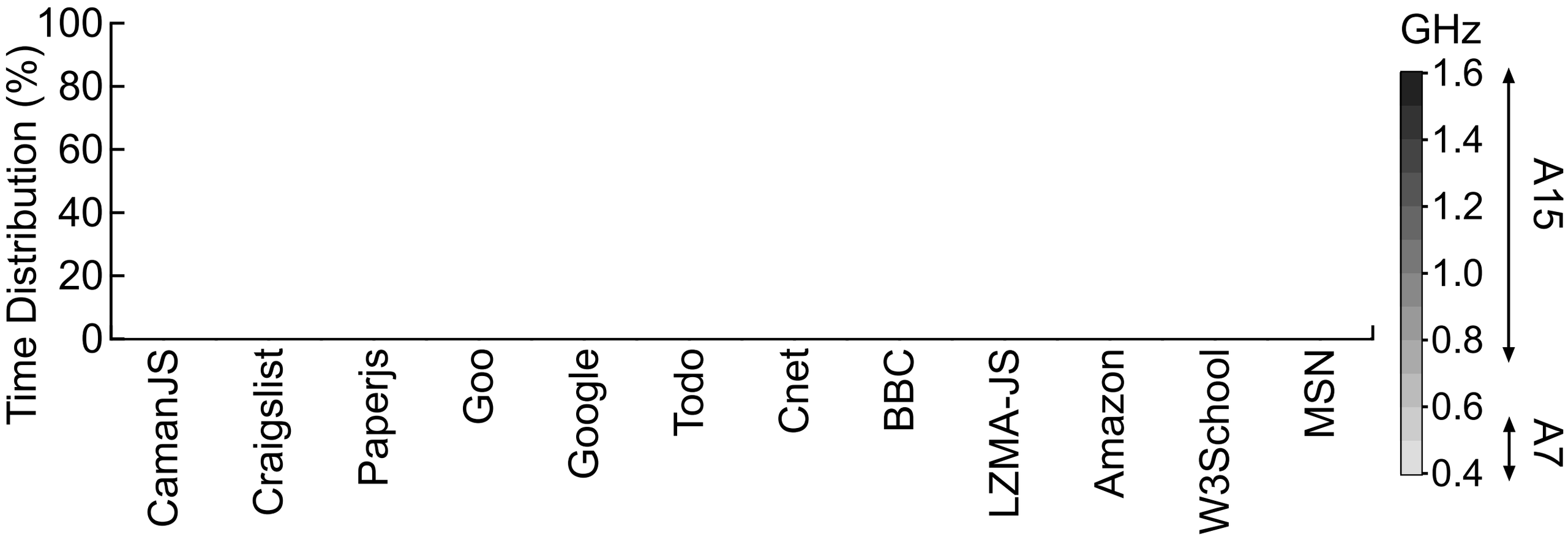


Evaluation Results

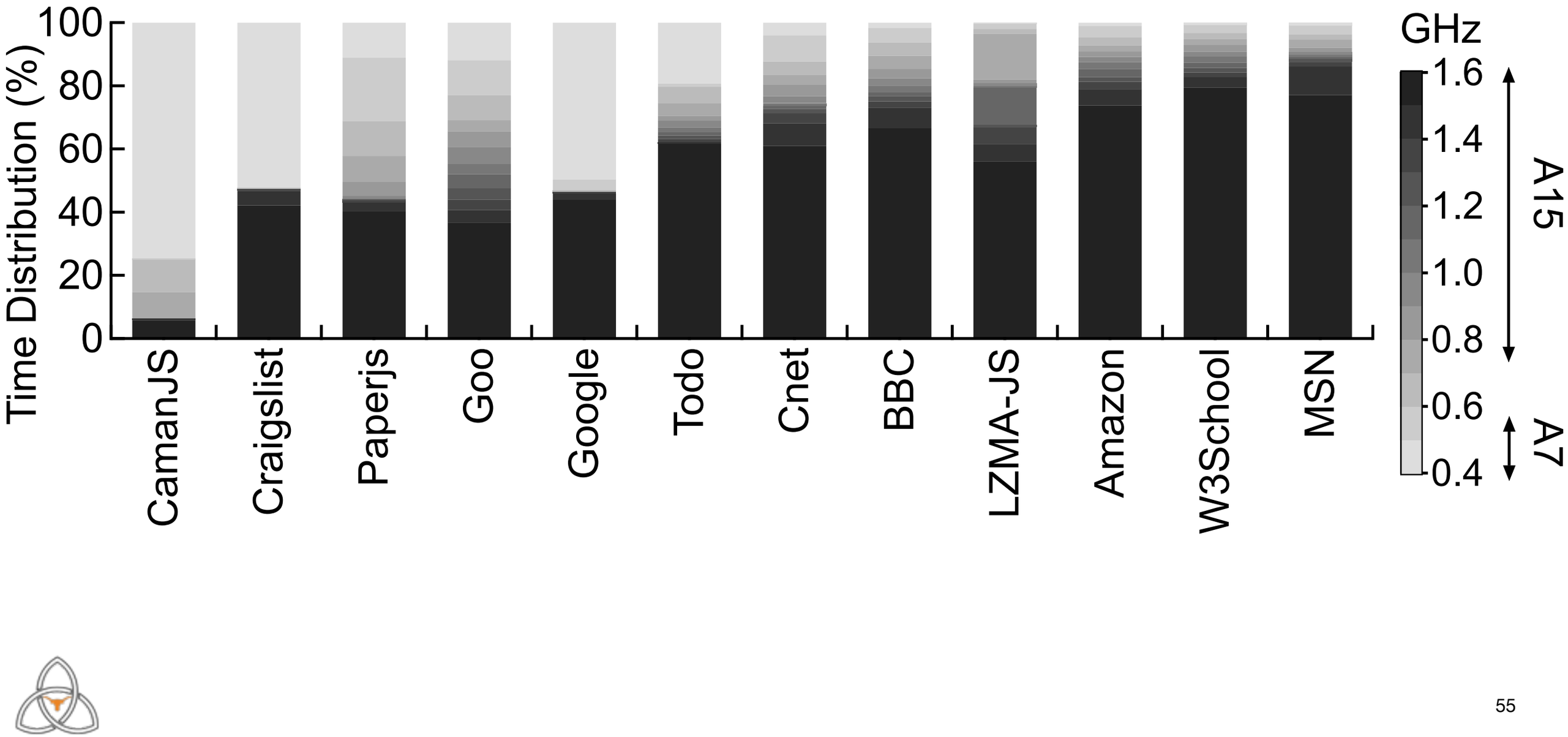


29.2% - 66.0% energy savings, **0.8%** more QoS violations

Architecture Configuration Distribution



Architecture Configuration Distribution



GreenWeb

Programming language support for balancing energy-efficiency and QoS in mobile Web computing

GreenWeb

Programming language support for balancing energy-efficiency and QoS in mobile Web computing

Abstraction

Express QoS constraints

GreenWeb

Programming language support for balancing energy-efficiency and QoS in mobile Web computing

Abstraction

Express QoS constraints

Runtime

Satisfy QoS specifications using energy saving techniques

GreenWeb

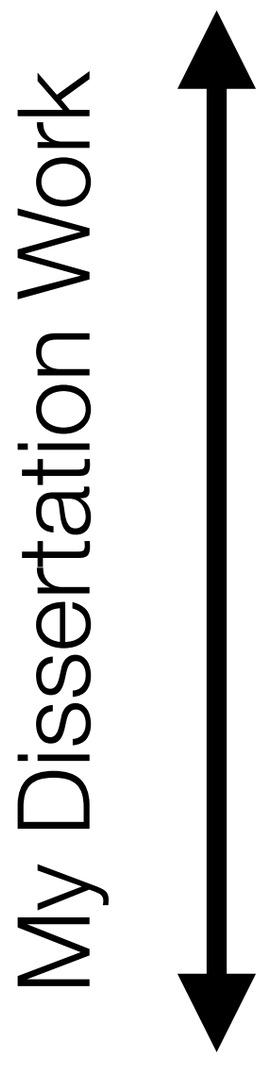
Programming language support for balancing energy-efficiency and QoS in mobile Web computing

Abstraction — Express QoS constraints

Runtime — Satisfy QoS specifications using energy saving techniques

Effect — Significant energy savings

My Approach



Application

GreenWeb

Language Extensions

Runtime

WebRT

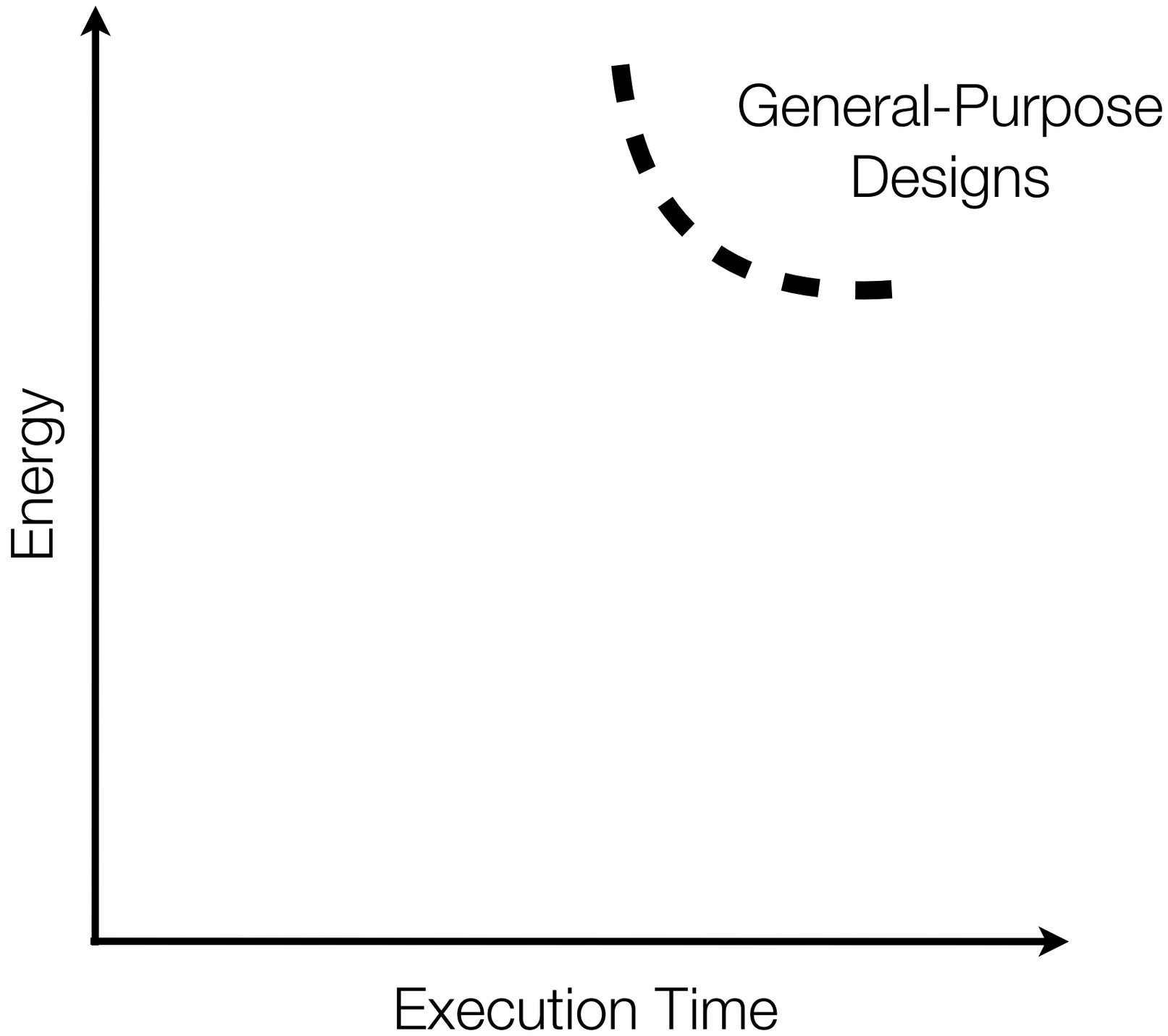
Energy-aware Web Runtime

Architecture

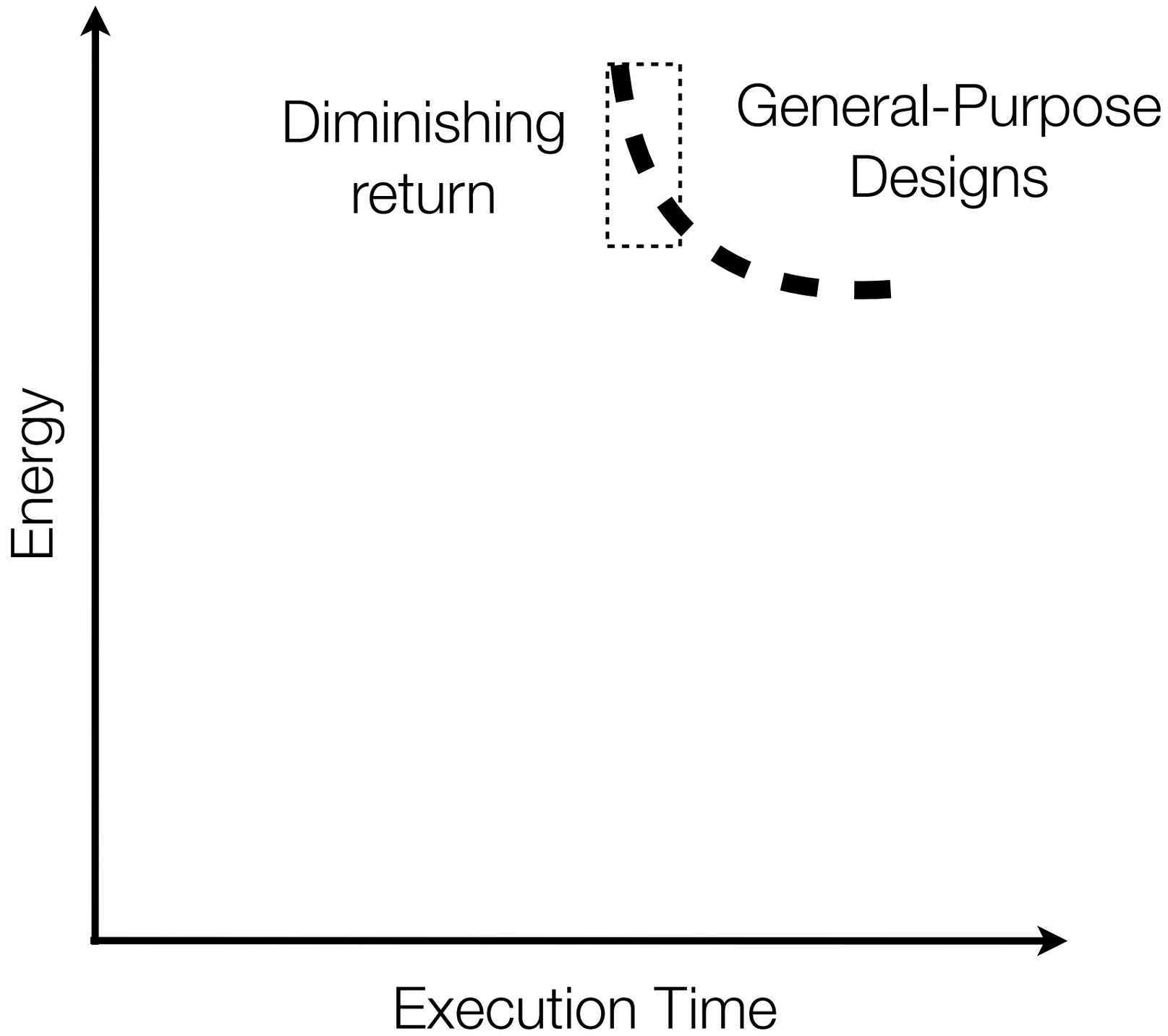
WebCore

Web-specific Architecture

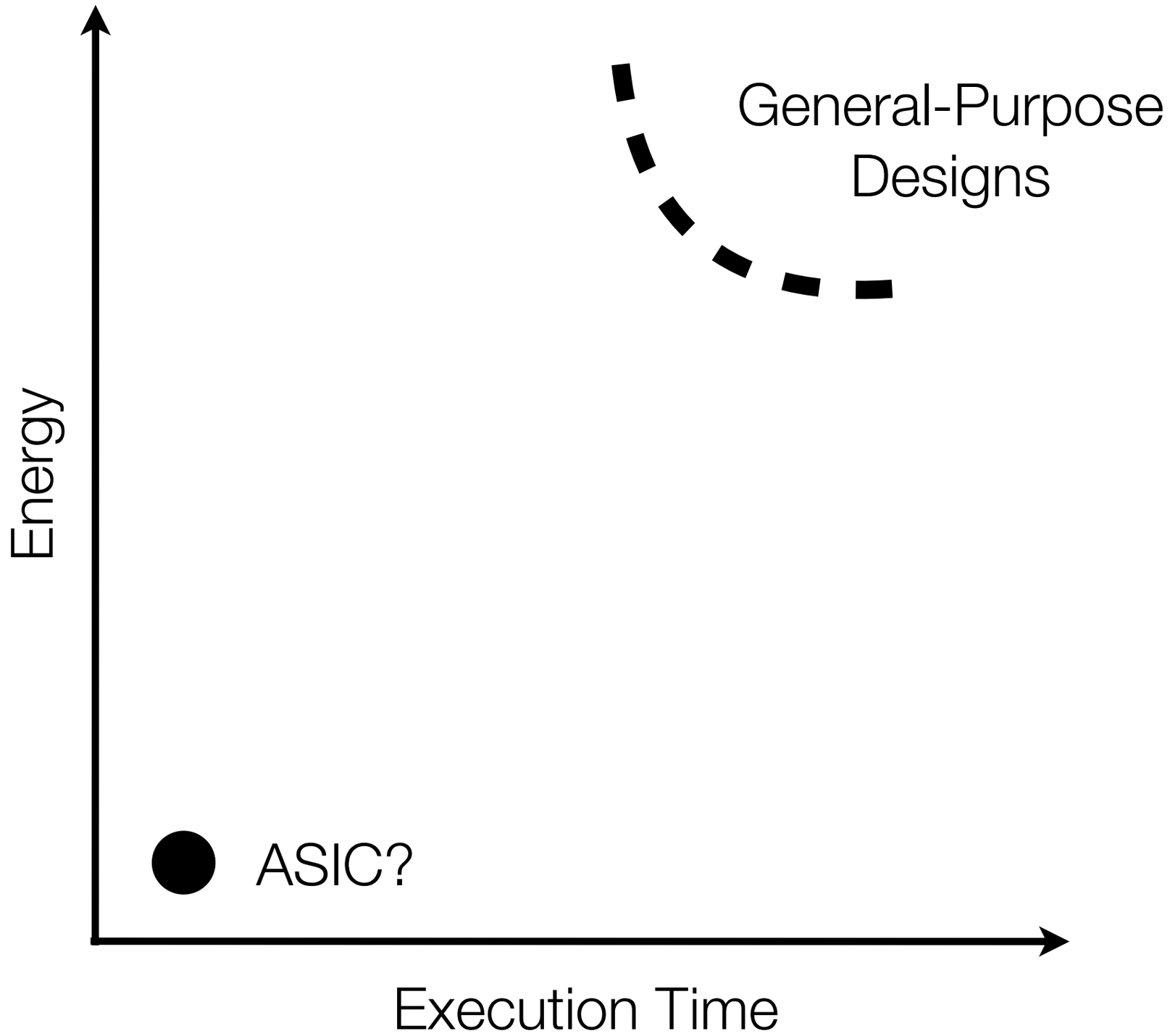
WebCore: a Web-Specific Mobile Architecture



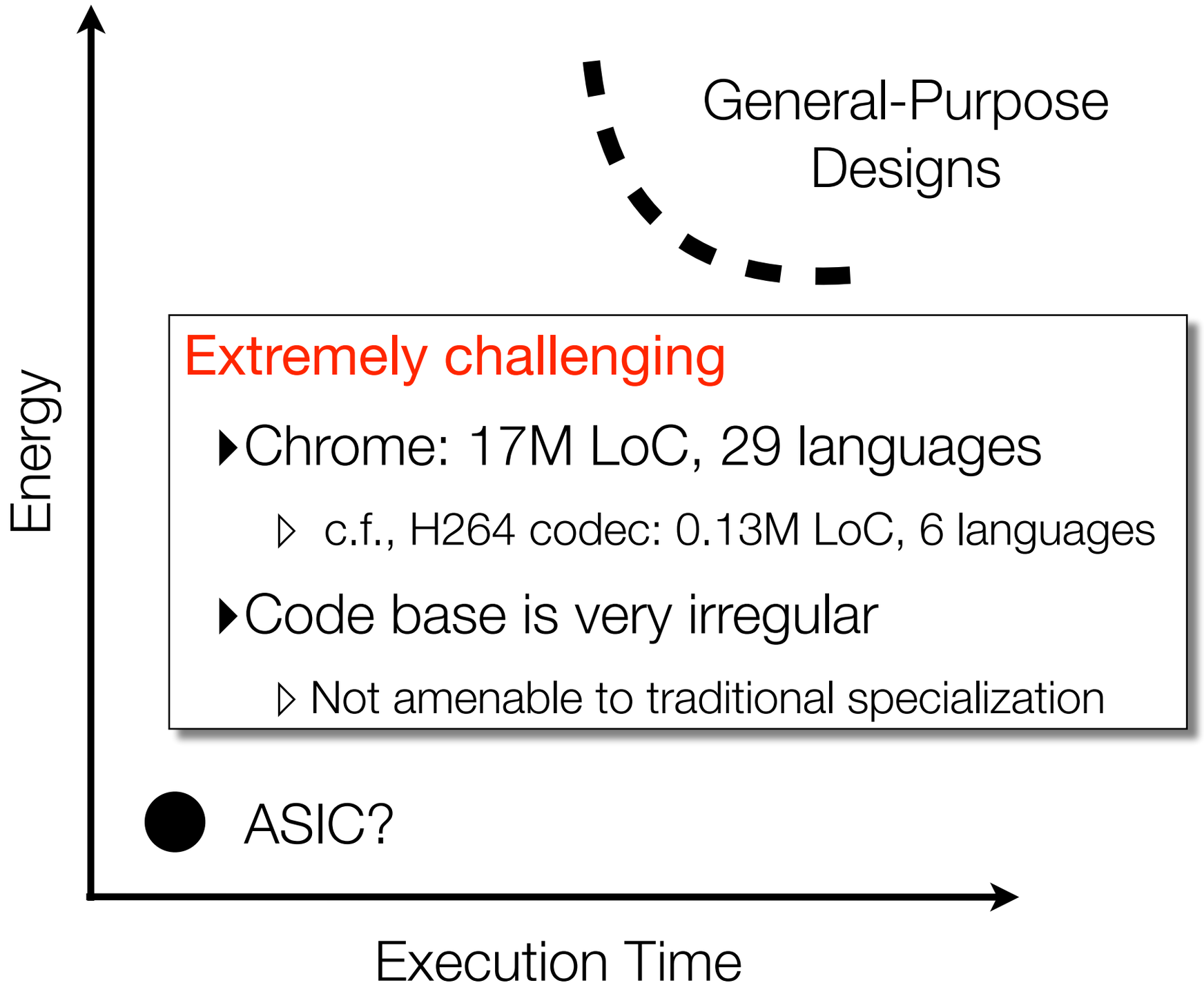
WebCore: a Web-Specific Mobile Architecture



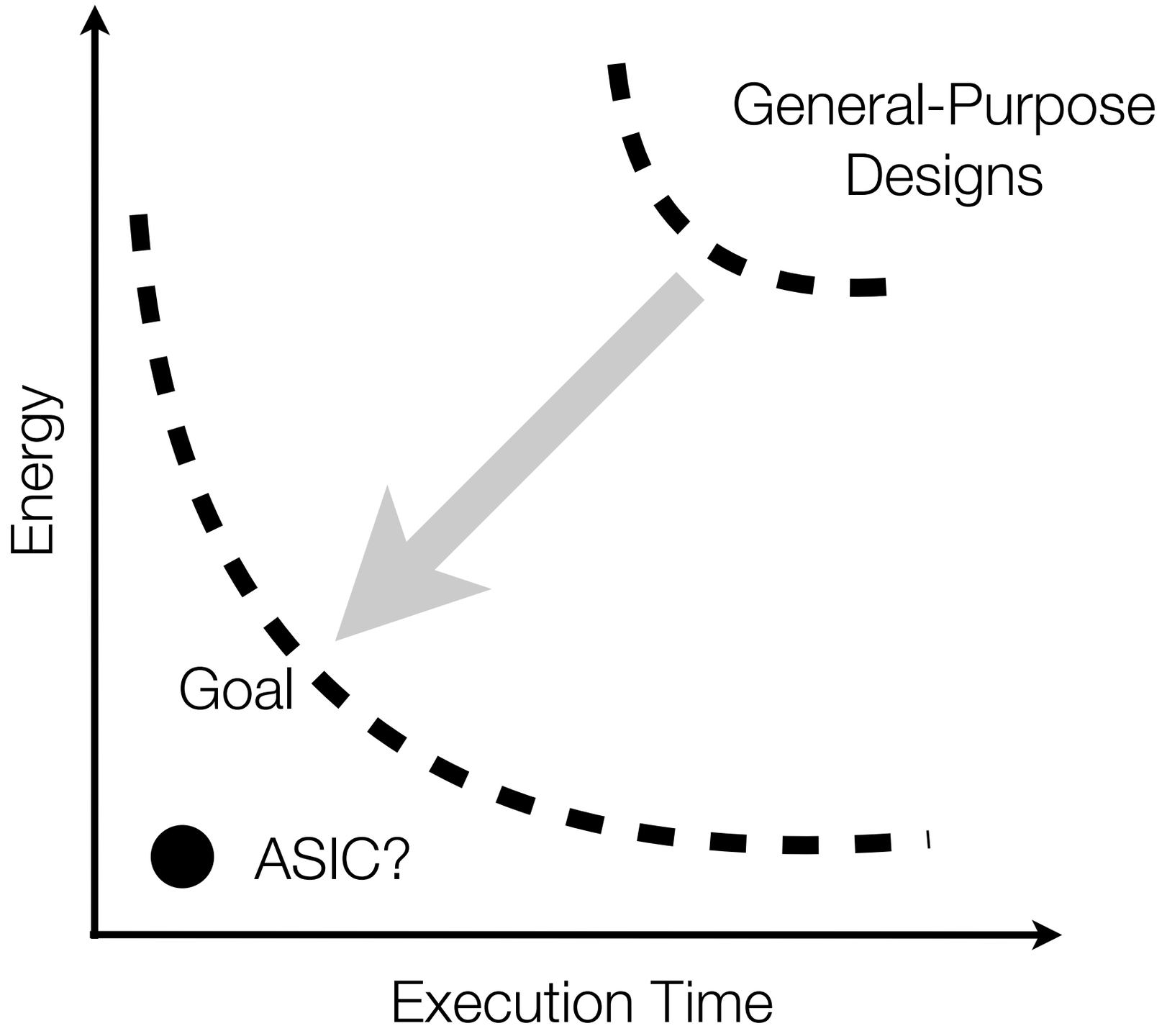
WebCore: a Web-Specific Mobile Architecture



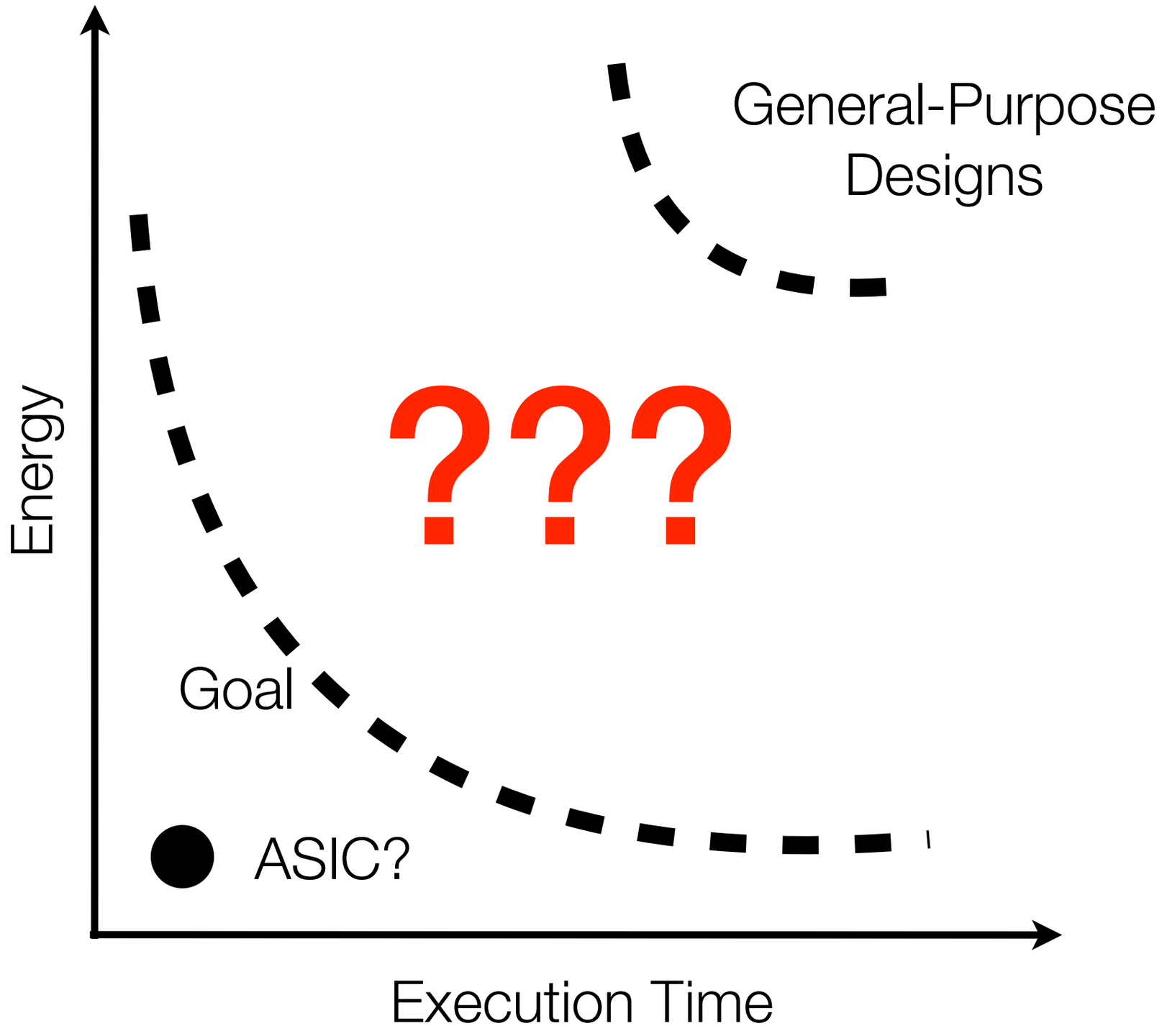
WebCore: a Web-Specific Mobile Architecture



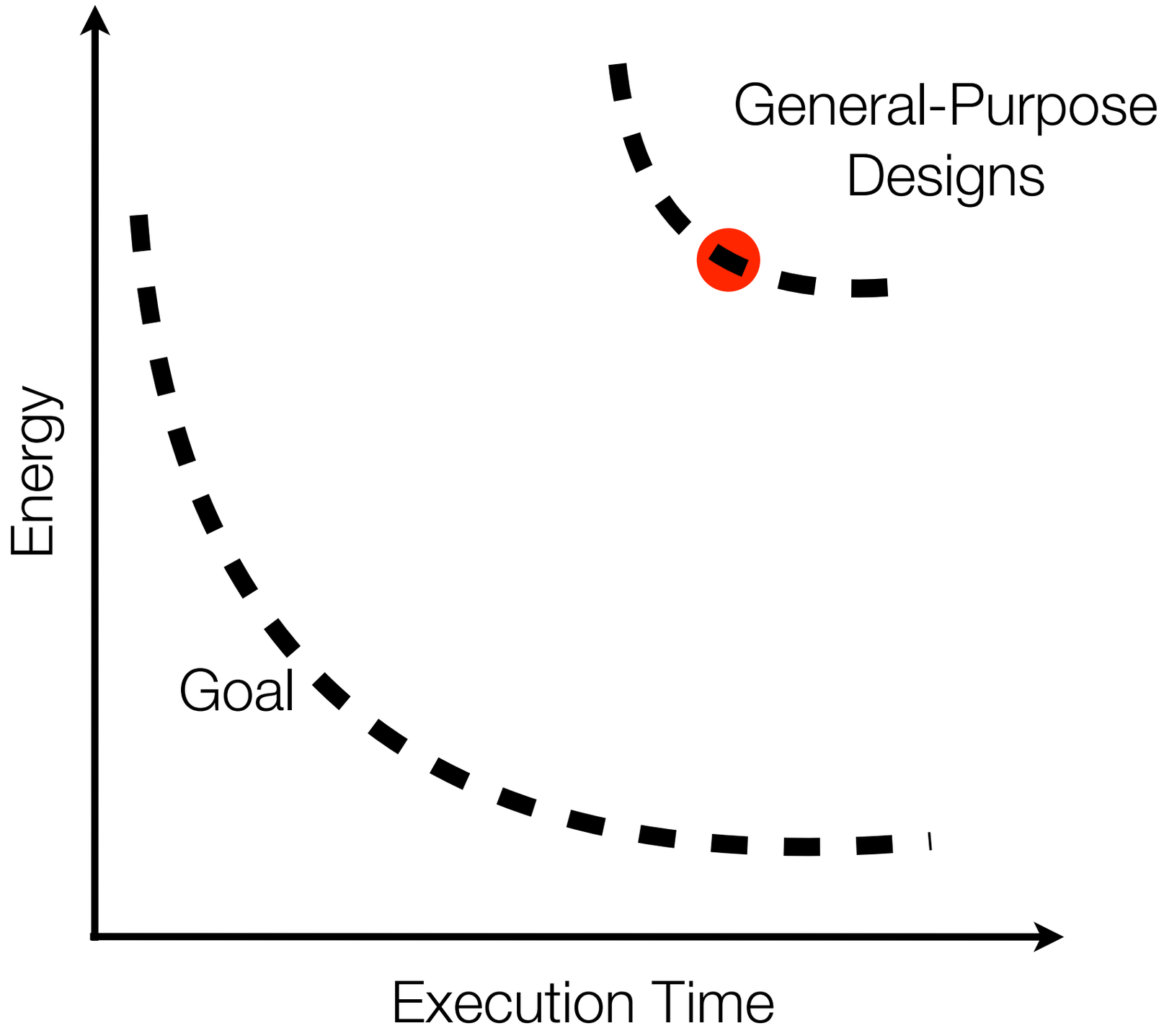
WebCore: a Web-Specific Mobile Architecture



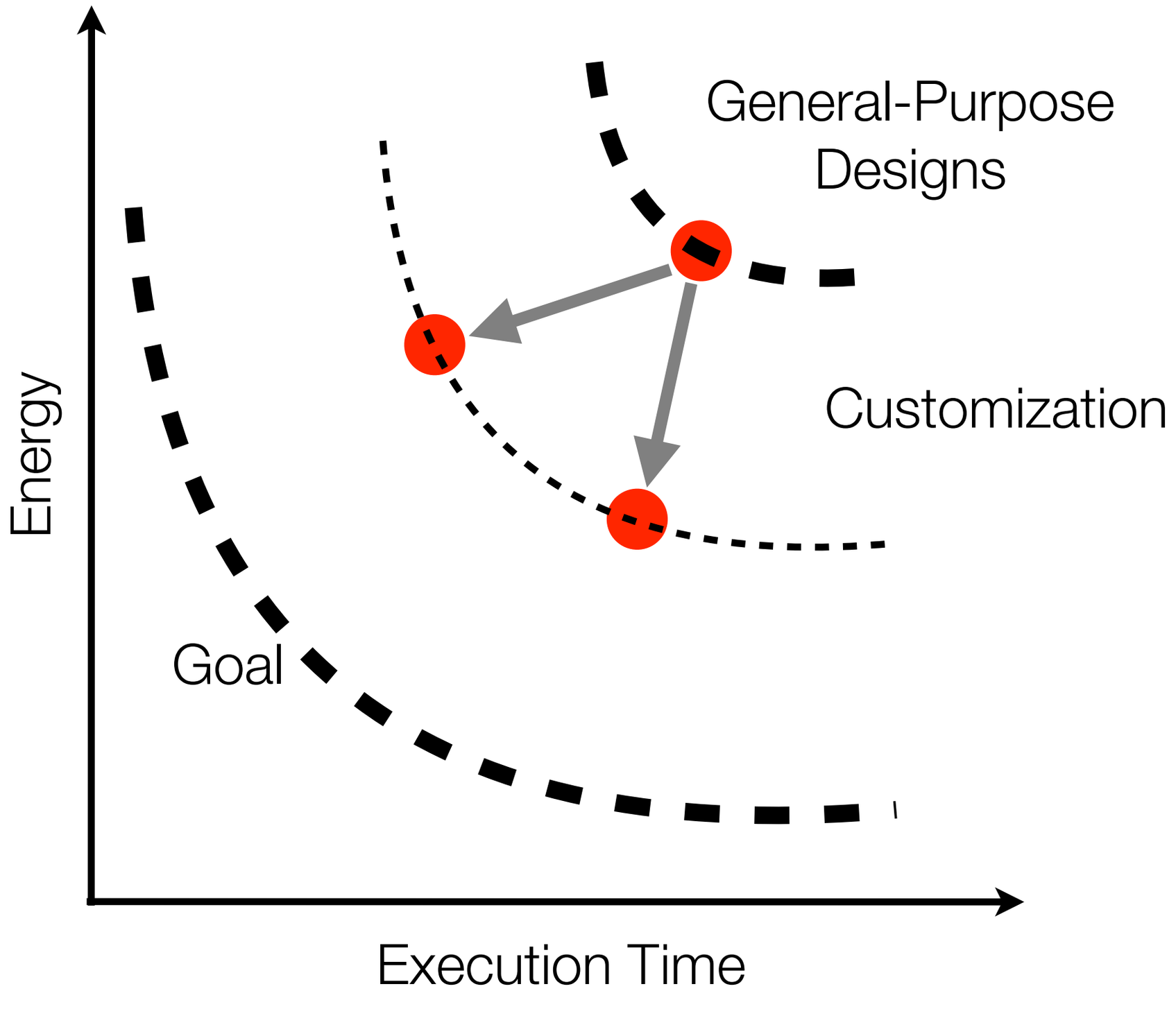
WebCore: a Web-Specific Mobile Architecture



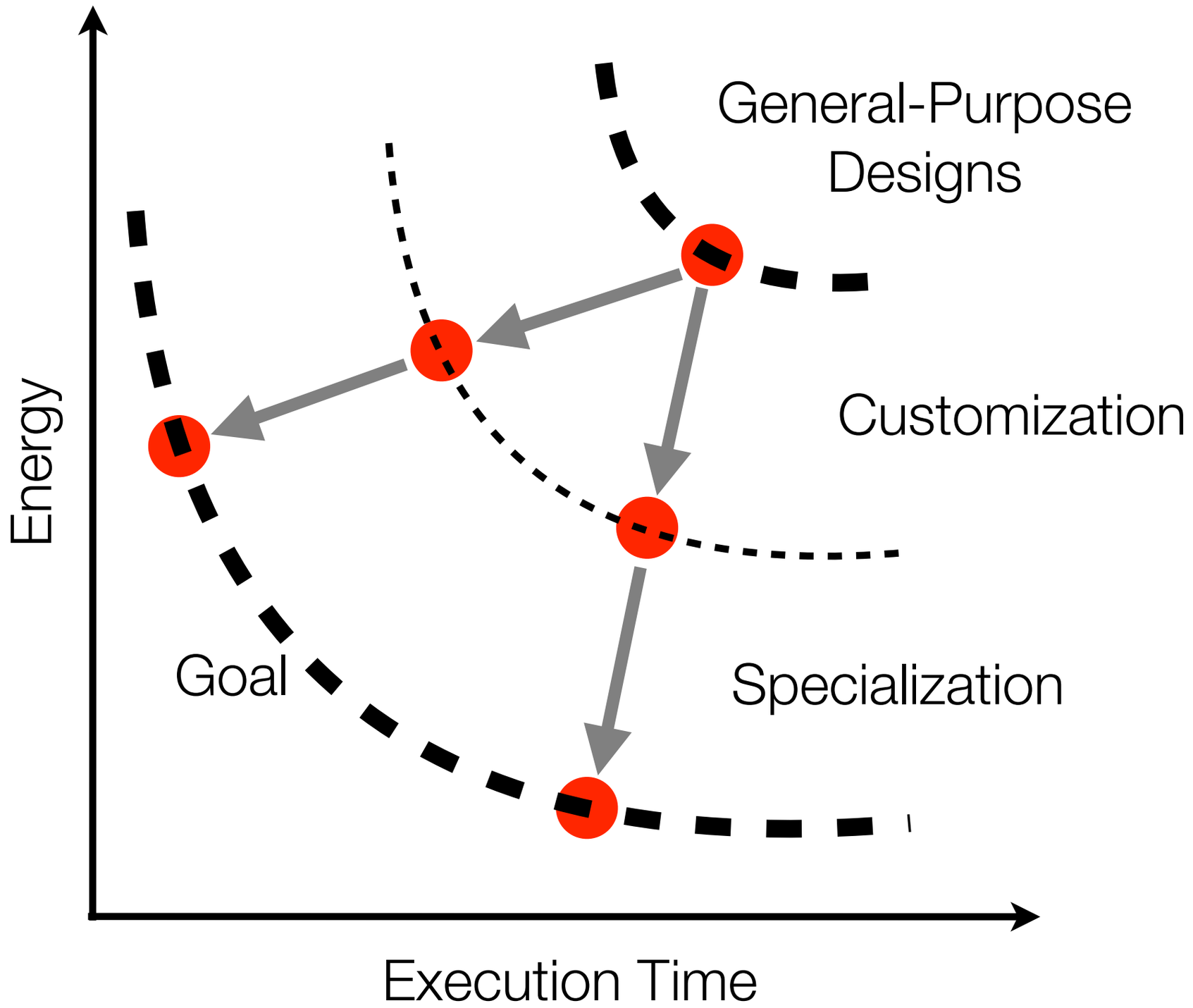
WebCore: a Web-Specific Mobile Architecture



WebCore: a Web-Specific Mobile Architecture

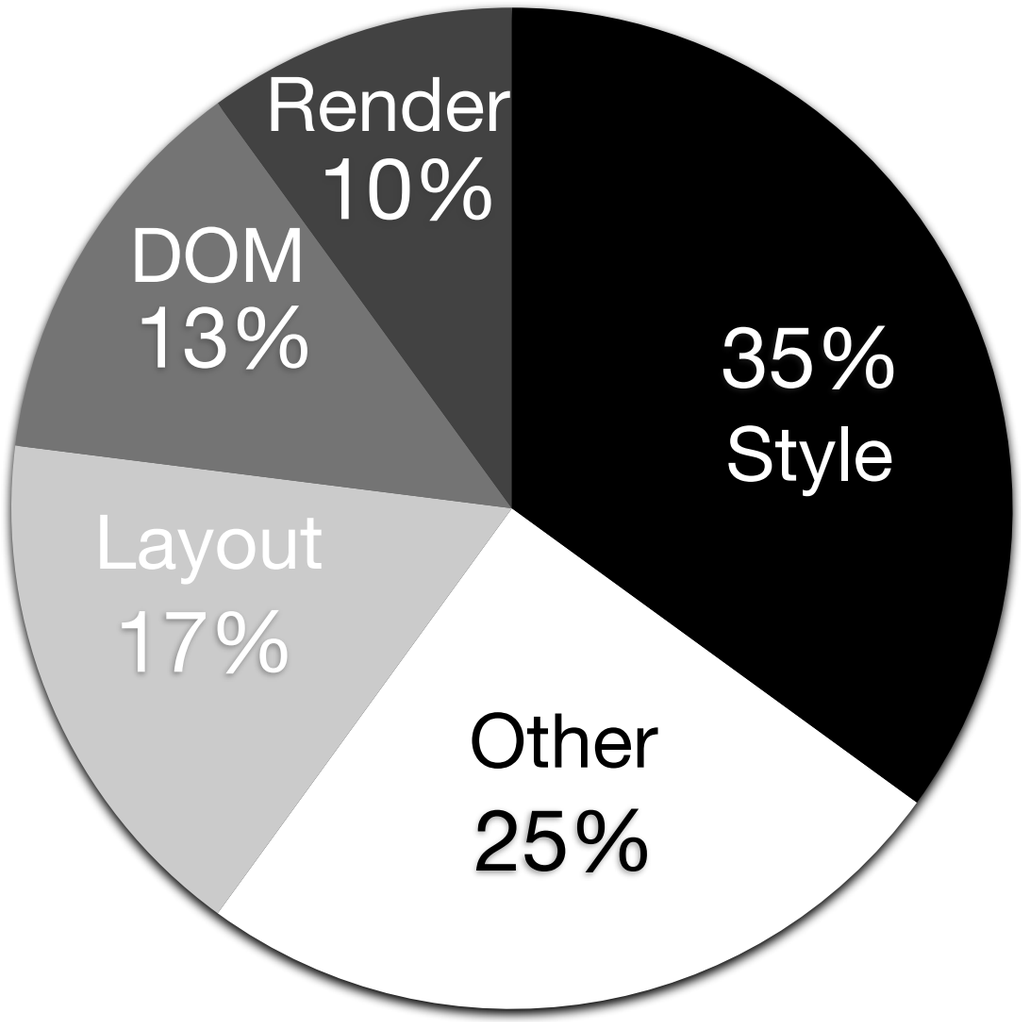


WebCore: a Web-Specific Mobile Architecture

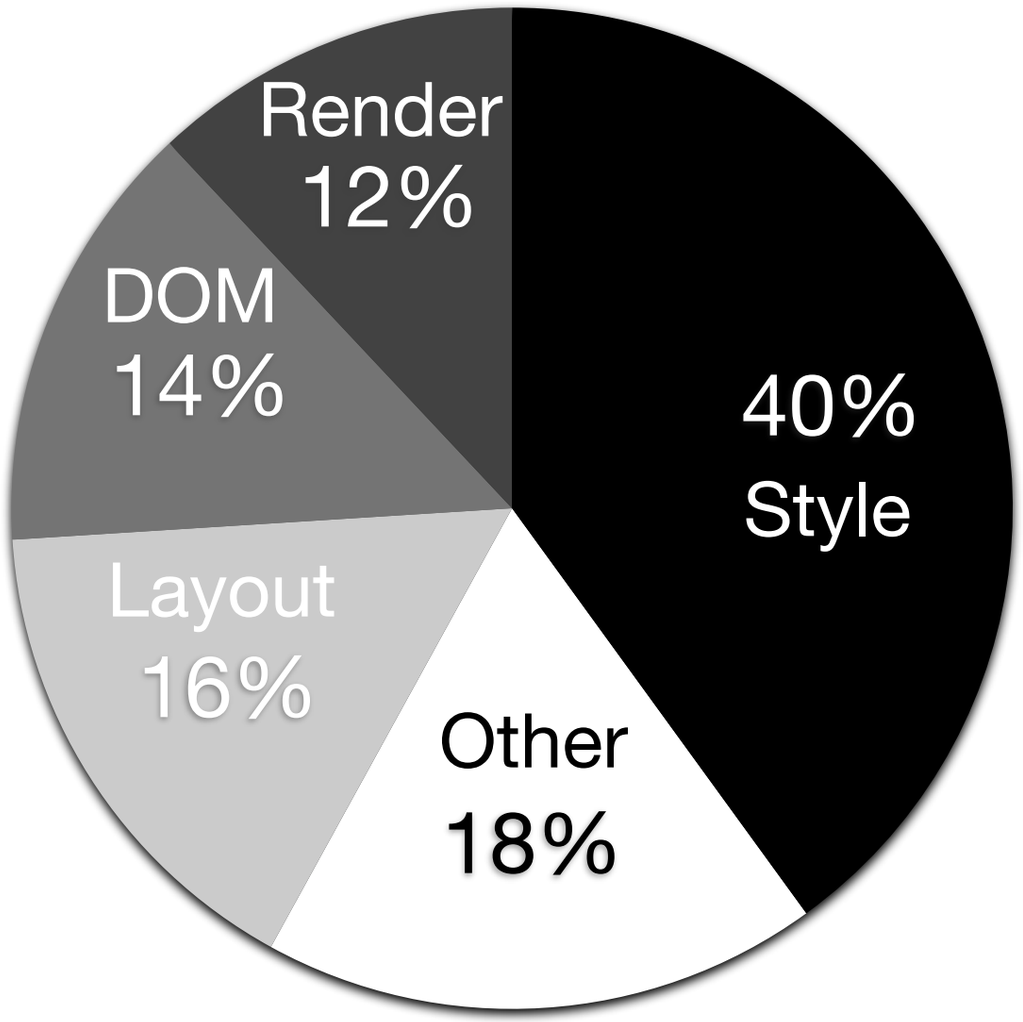


Specialization Target: Style Resolution Kernel

Specialization Target: Style Resolution Kernel

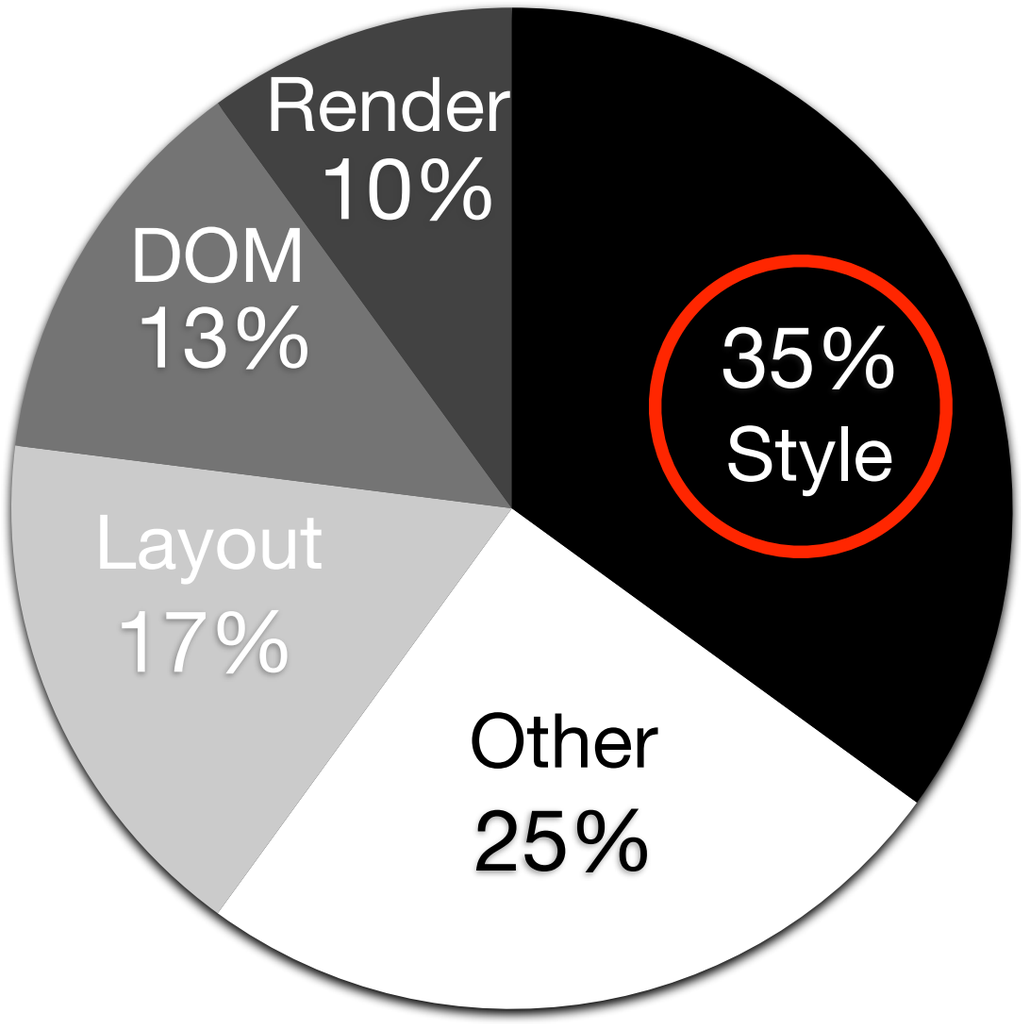


Execution time breakdown

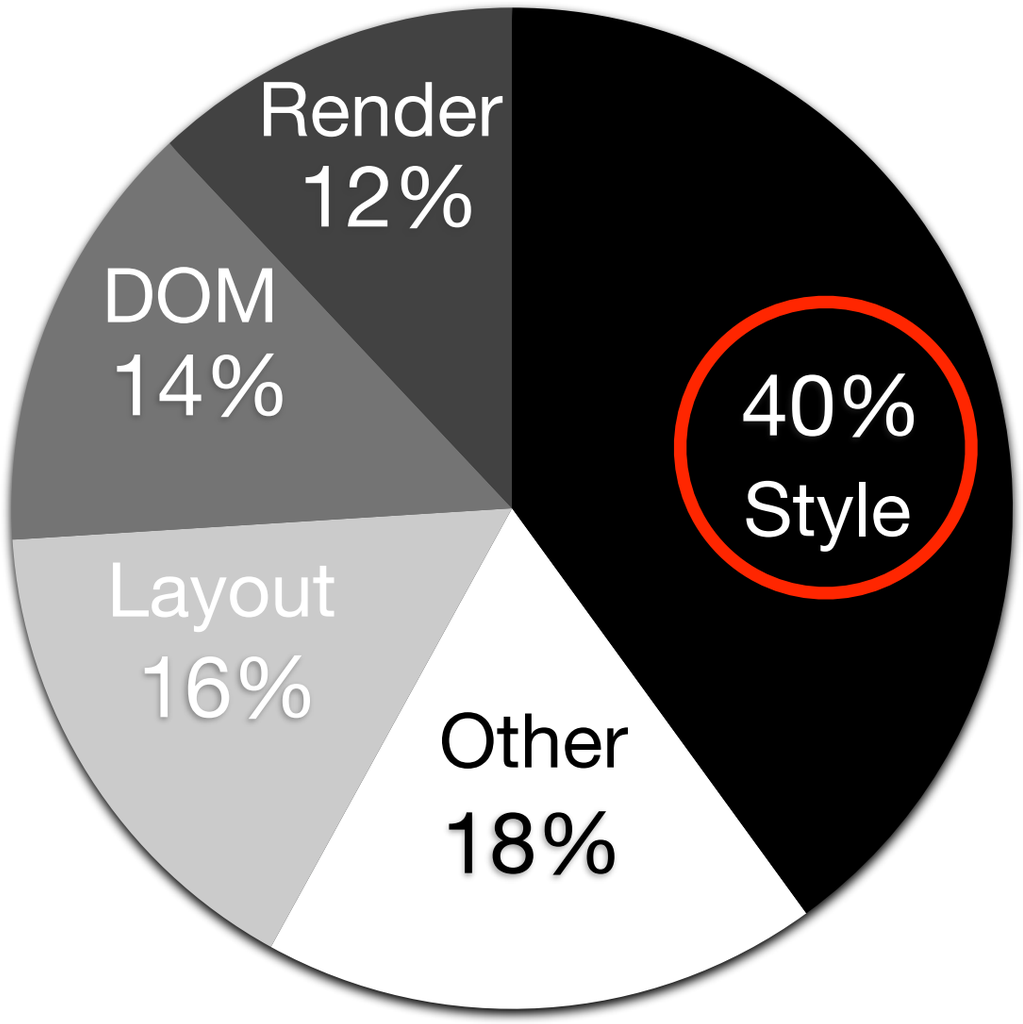


Energy breakdown

Specialization Target: Style Resolution Kernel



Execution time breakdown



Energy breakdown

Specialization Target: Style Resolution Kernel

```
for (each rule in matchedRules) {
  for (each property in rule) {
    switch (property.id) {
      case Font:
        Style[Font] = Handler(property.value, DOMNode);
        break;
      case N: ...}}}
```

Specialization Target: Style Resolution Kernel

```
for (each rule in matchedRules) {
  for (each property in rule) {
    switch (property.id) {
      case Font:
        Style[Font] = Handler(property.value, DOMNode);
        break;
      case N: ...}}}
```

Specialization Target: Style Resolution Kernel

```
for (each rule in matchedRules) {
  for (each property in rule) {
    switch (property.id) {
      case Font:
        Style[Font] = Handler(property.value, DOMNode);
        break;
      case N: ...}}}
```

Rule-level Parallelism (RLP)

Specialization Target: Style Resolution Kernel

```
for (each rule in matchedRules) {
  for (each property in rule) {
    switch (property.id) {
      case Font:
        Style[Font] = Handler(property.value, DOMNode);
        break;
      case N: ...}}}
```

Rule-level Parallelism (RLP)

Specialization Target: Style Resolution Kernel

```
for (each rule in matchedRules) {
  for (each property in rule) {
    switch (property.id) {
      case Font:
        Style[Font] = Handler(property.value, DOMNode);
        break;
      case N: ...}}}
```

Rule-level Parallelism (RLP)

Property-level Parallelism (PLP)

Specialization Target: Style Resolution Kernel

```
for (each rule in matchedRules) {
  for (each property in rule) {
    switch (property.id) {
      case Font:
        Style[Font] = Handler(property.value, DOMNode);
        break;
      case N: ...}}}
```

Rule-level Parallelism (RLP)

Property-level Parallelism (PLP)

- Exploiting the parallelism to increase the arithmetic intensity
- Move operands closer to operations to sustain the computations

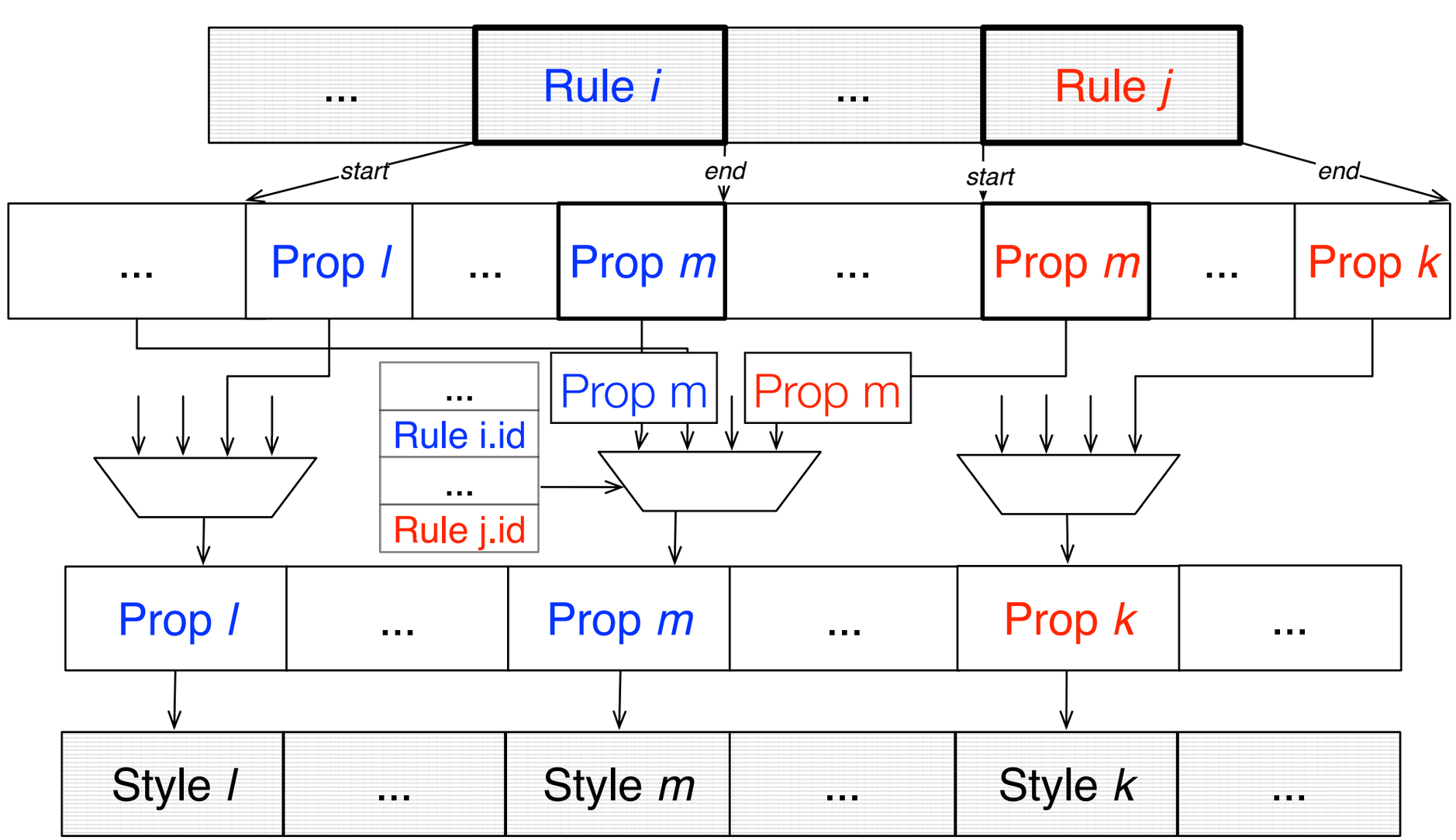
Style Resolution Unit

Input Scratchpad

Conflict Resolution

Compute Lanes

Output Scratchpad



Evaluation Results

Evaluation Results

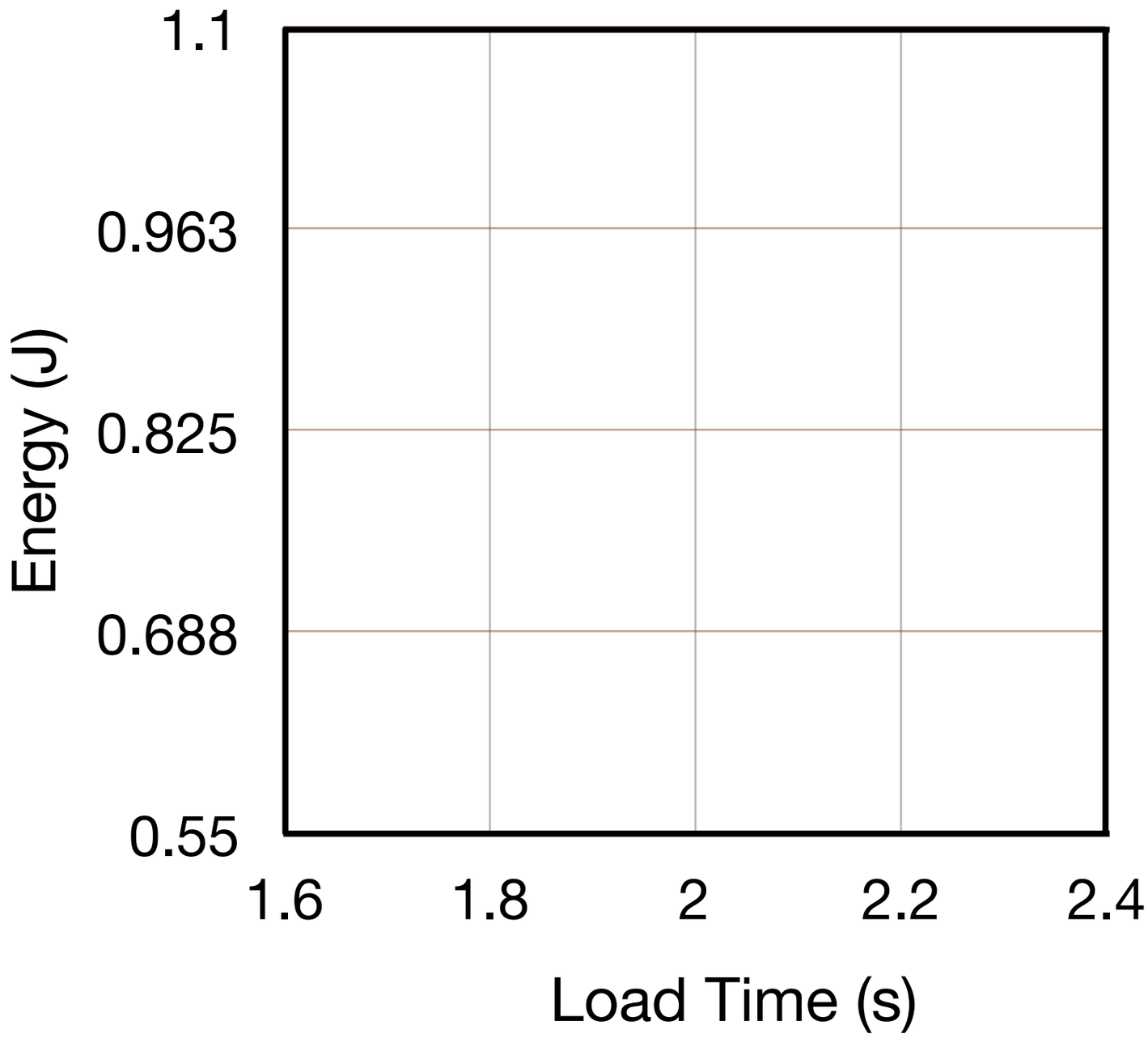
- Fully synthesized using Synopsys 28 nm toolchain

Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2

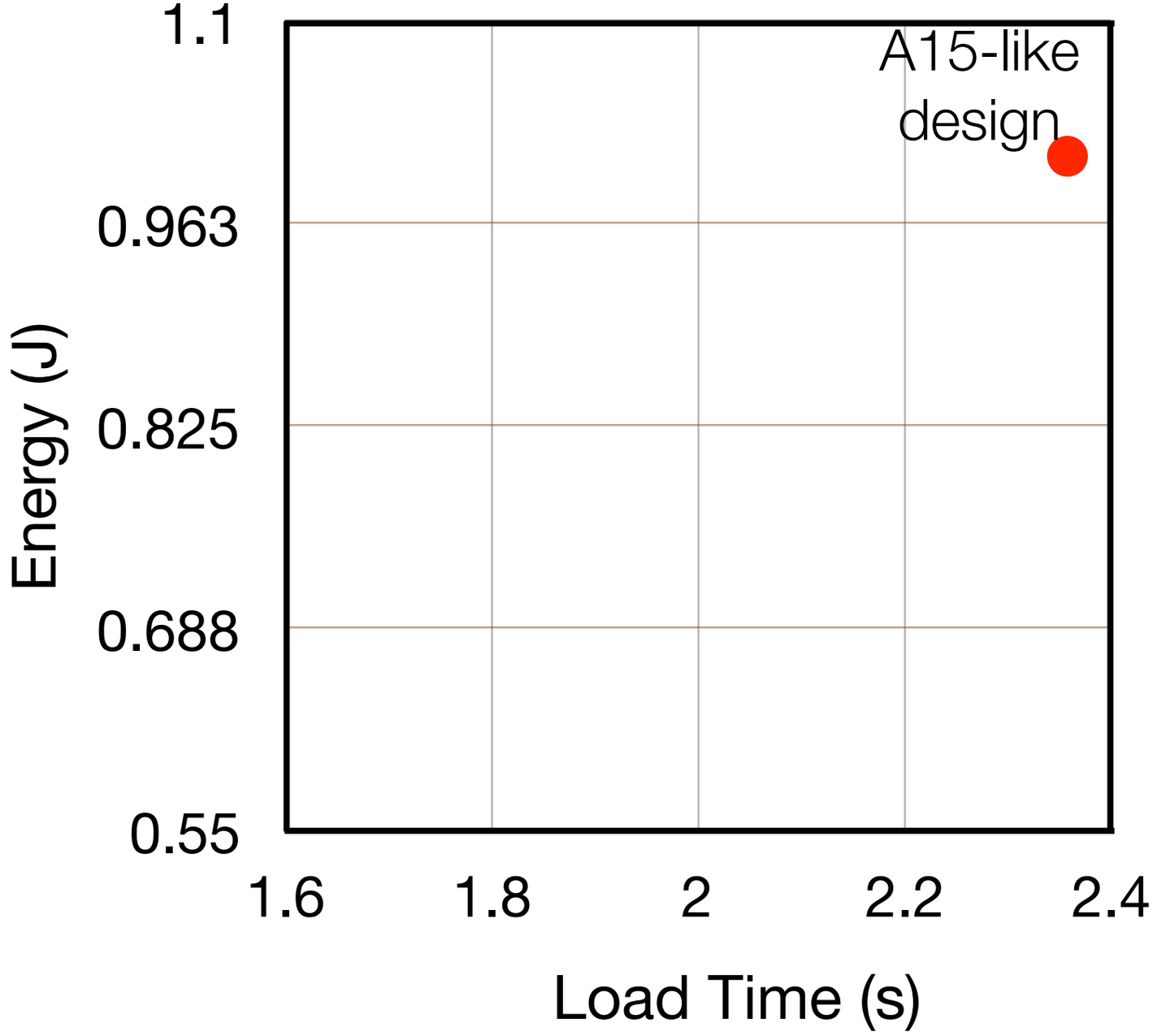
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2



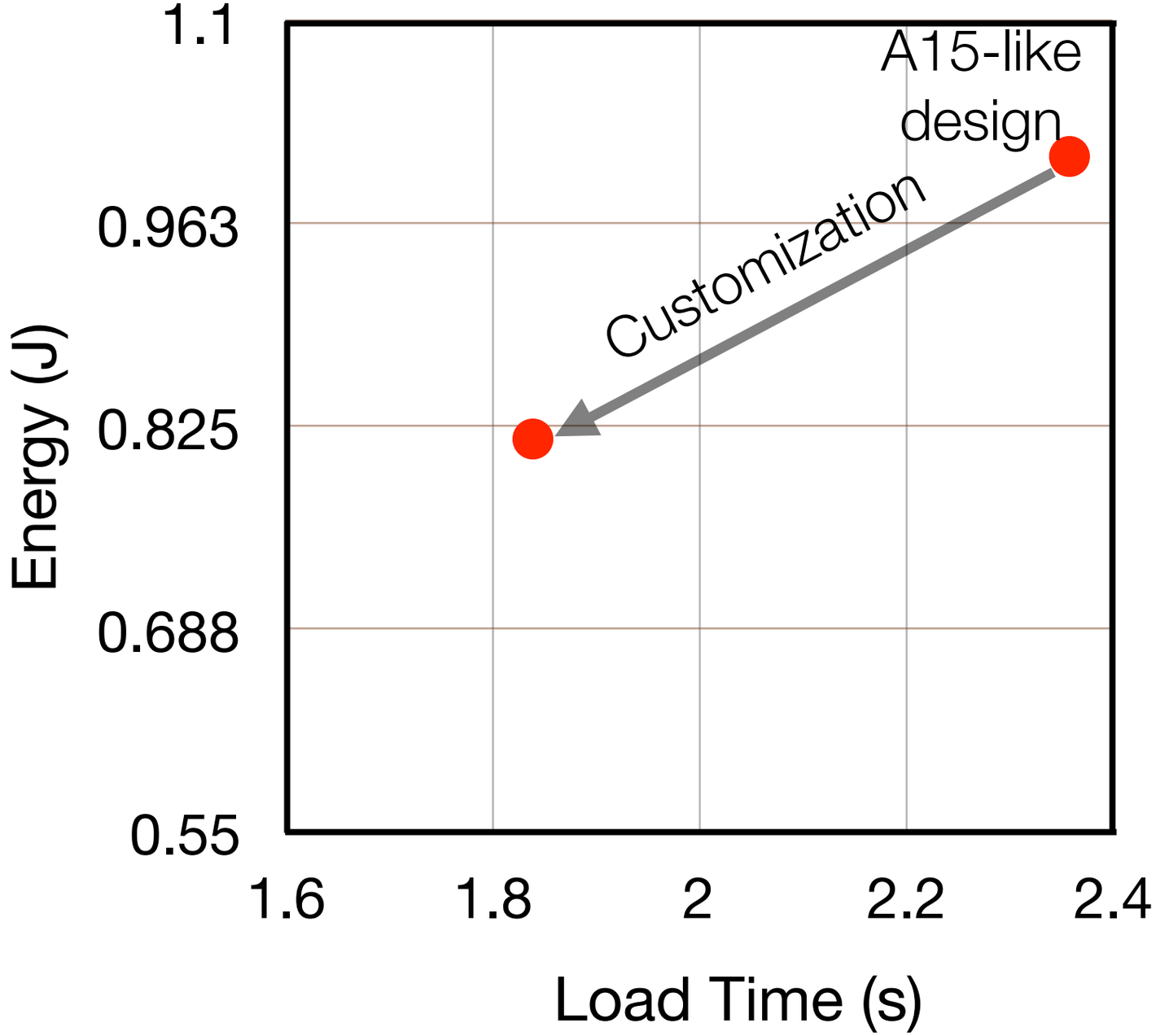
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2



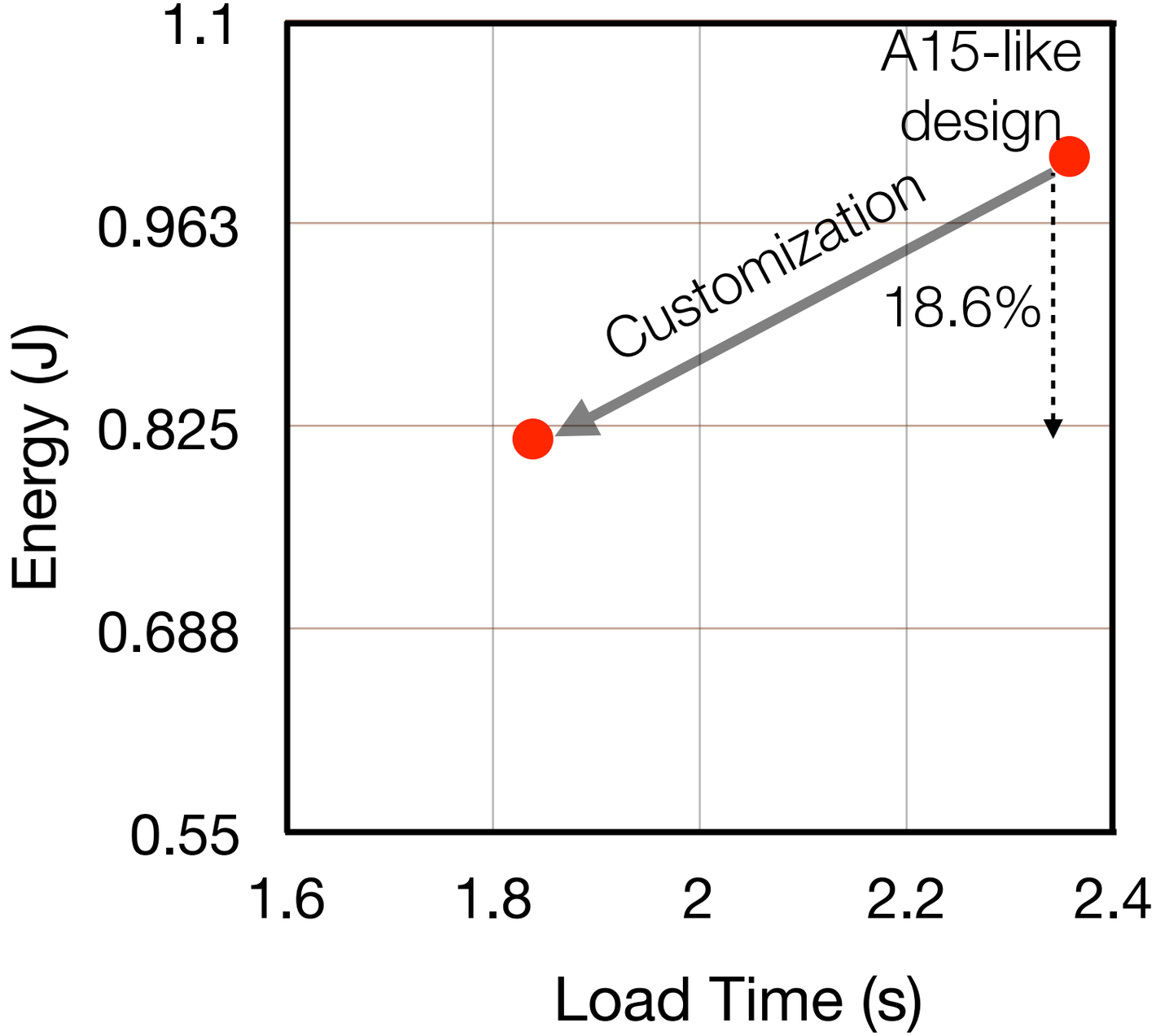
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2



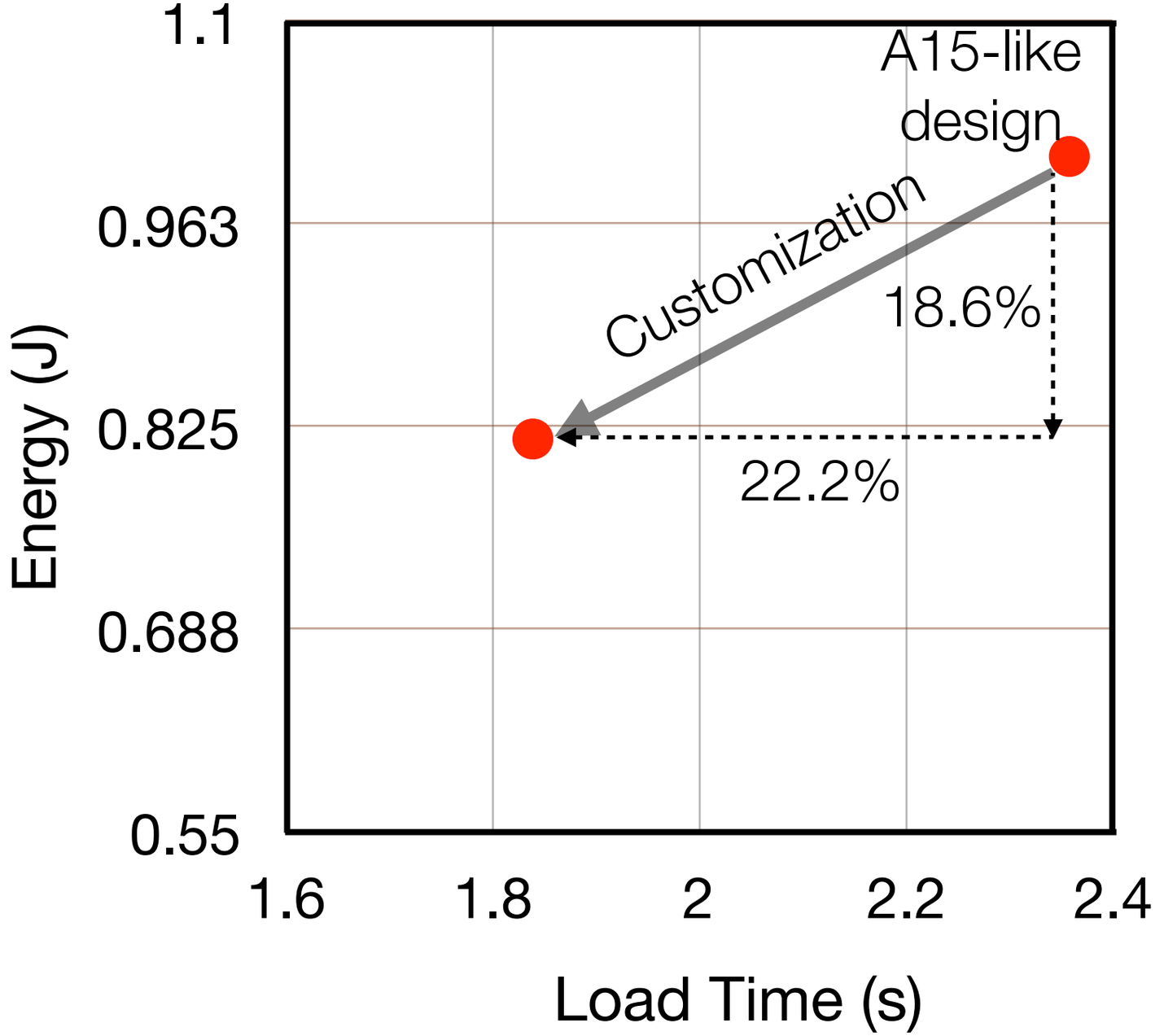
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2



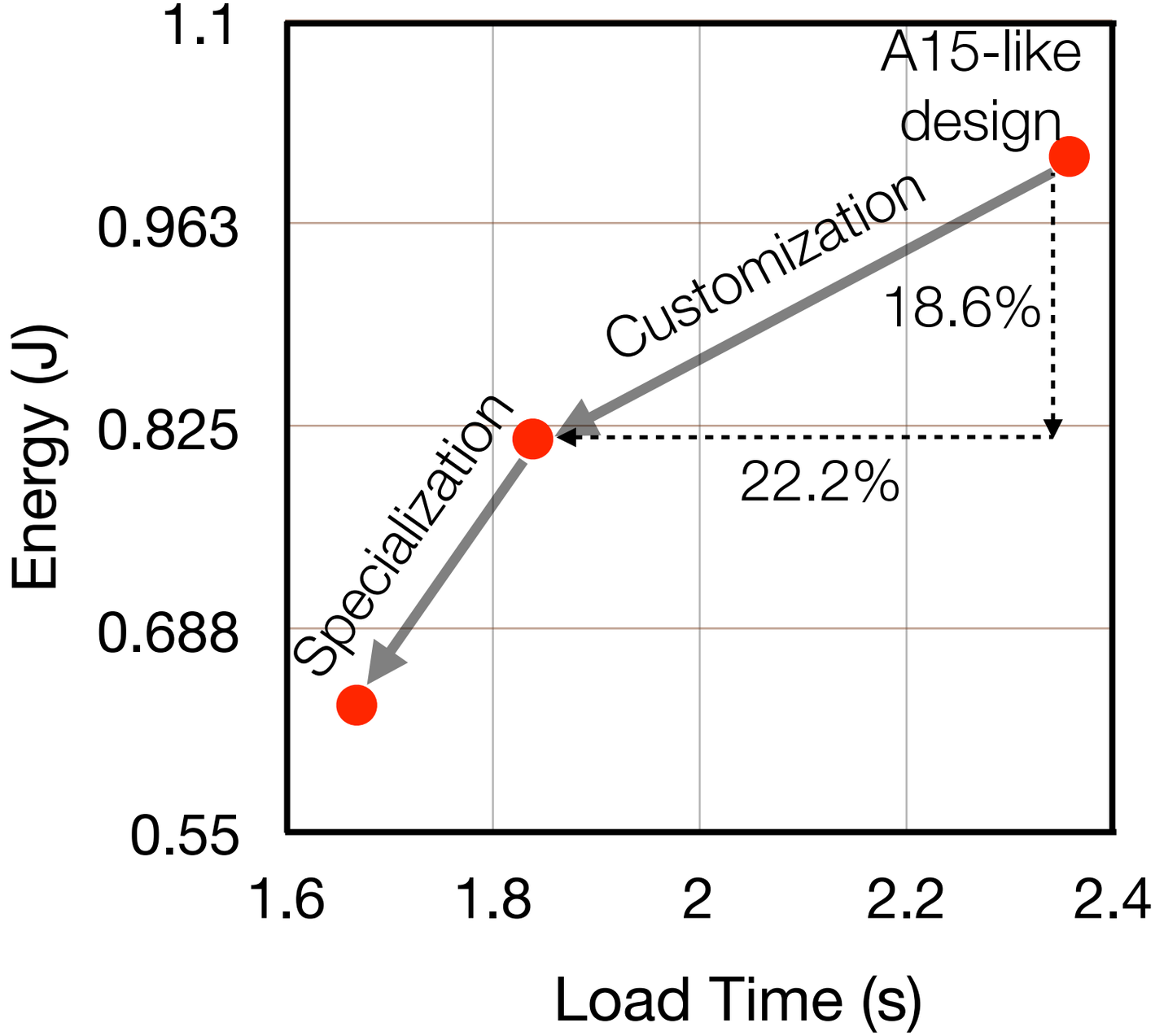
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2



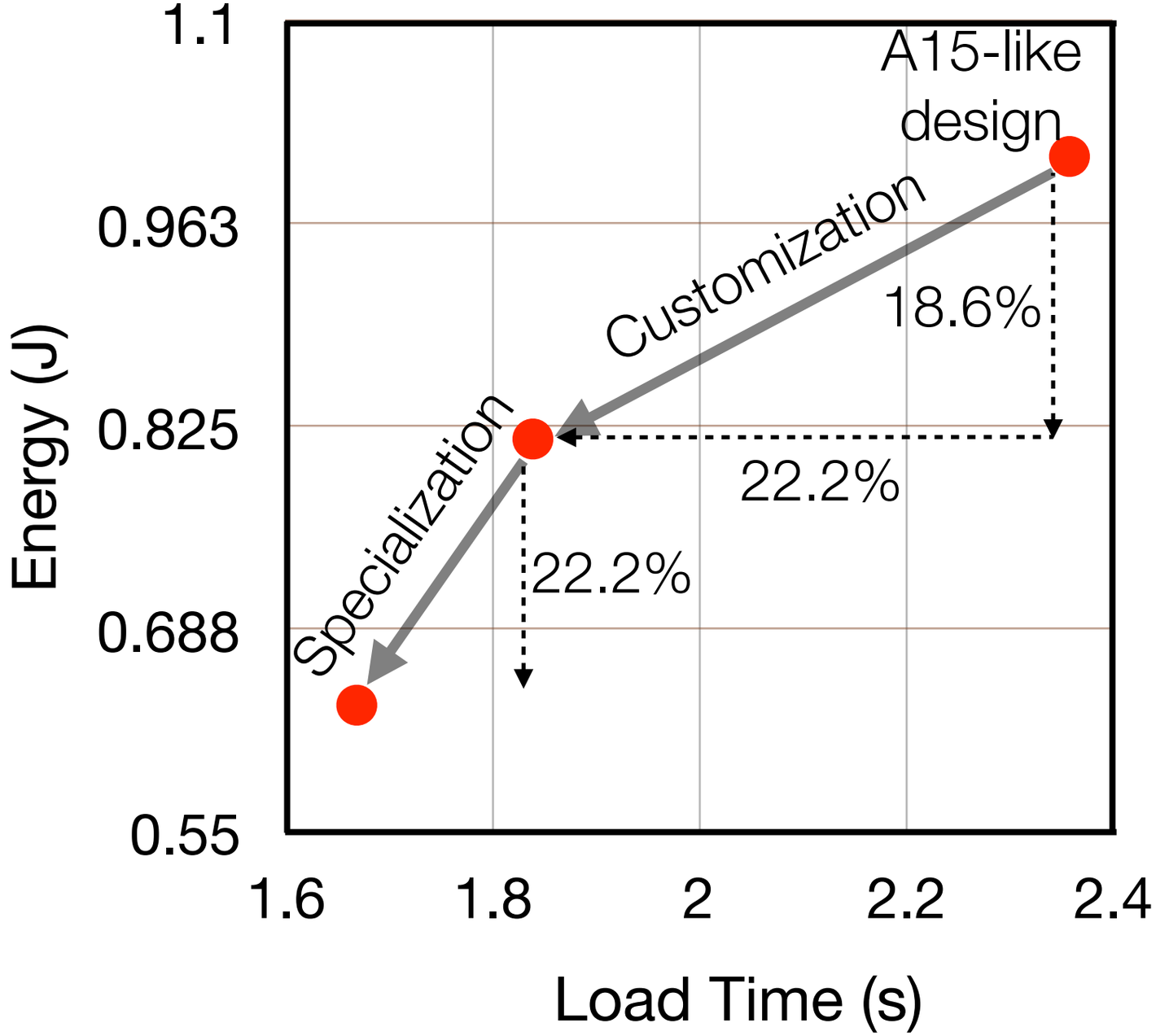
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2



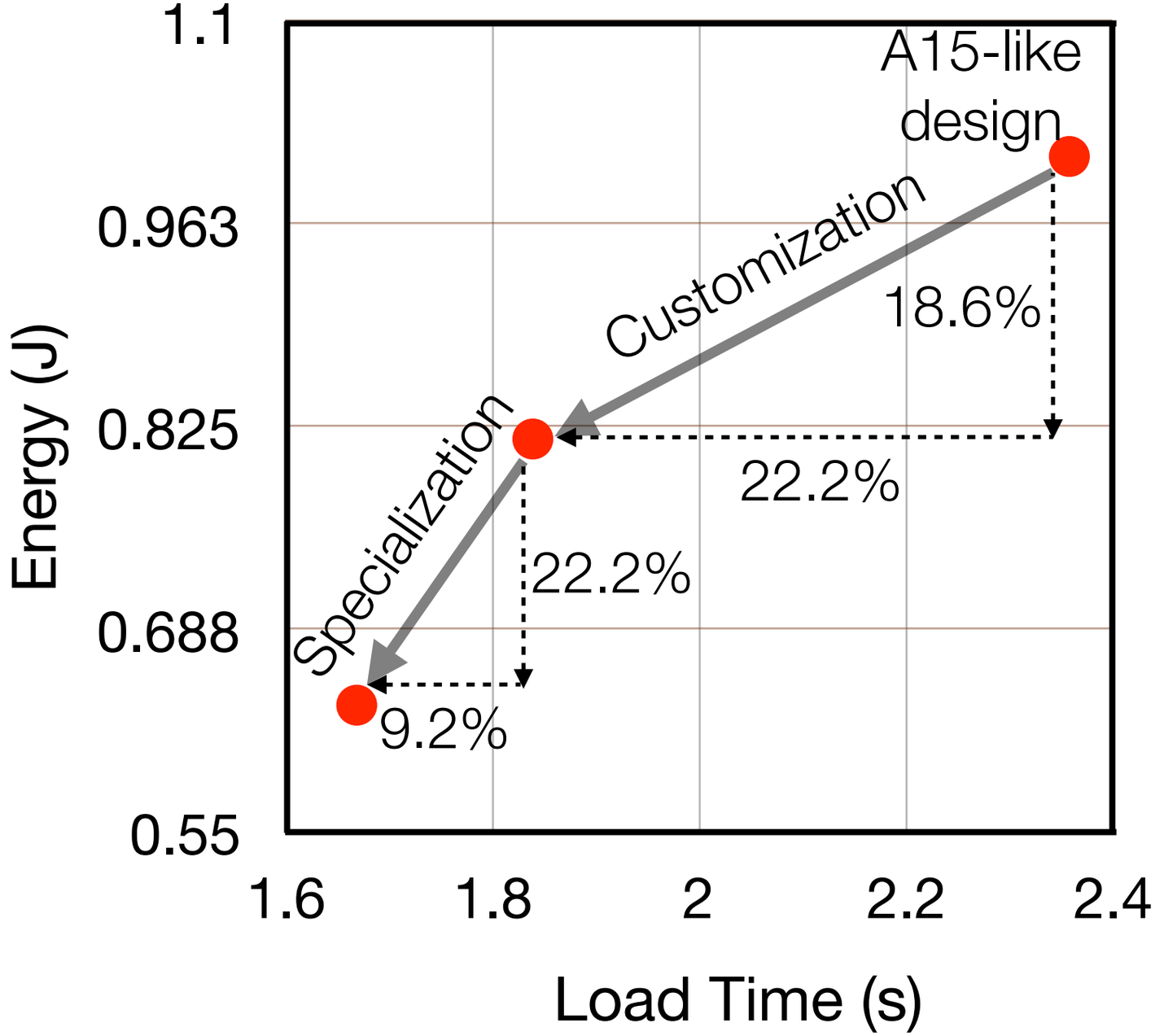
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2



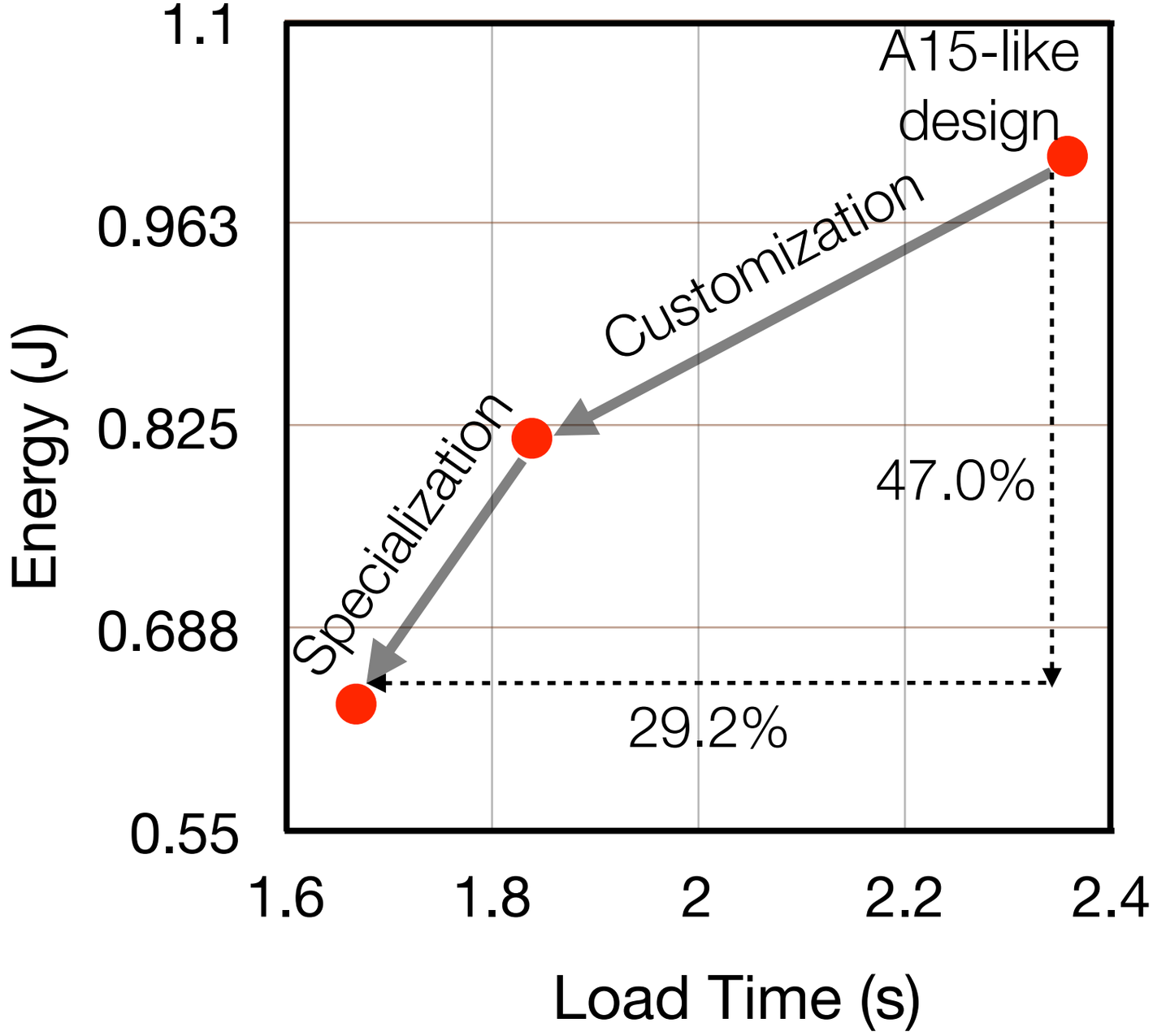
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2



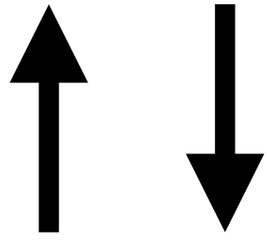
Evaluation Results

- Fully synthesized using Synopsys 28 nm toolchain
- Cost of specialization: 0.59 mm^2 area overhead
  - SoC die area is 122 mm^2 in Samsung Galaxy S4
  - A15s' area: 19 mm^2

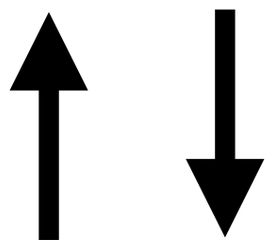


Retrospective: Three Principles Learnt

Application



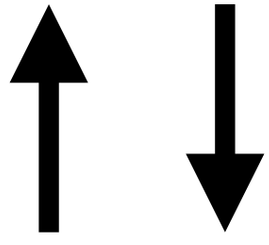
Runtime



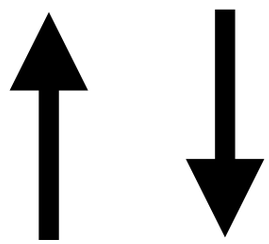
Architecture

Retrospective: Three Principles Learnt

Application



Runtime

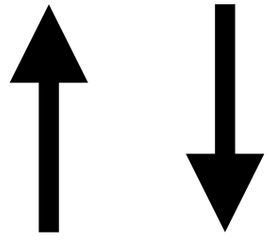


Architecture

- General-purpose vs. Specialization

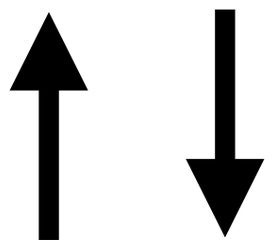
Retrospective: Three Principles Learnt

Application



Runtime

- Exploiting Application Diversity



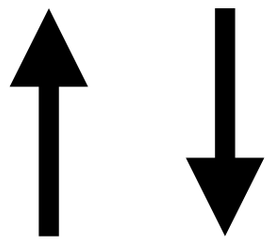
Architecture

- General-purpose vs. Specialization

Retrospective: Three Principles Learnt

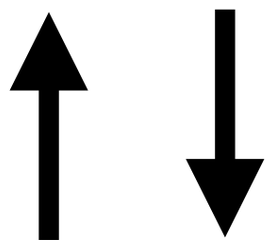
Application

- Empowering Web Developers



Runtime

- Exploiting Application Diversity



Architecture

- General-purpose vs. Specialization

The Web Evolution

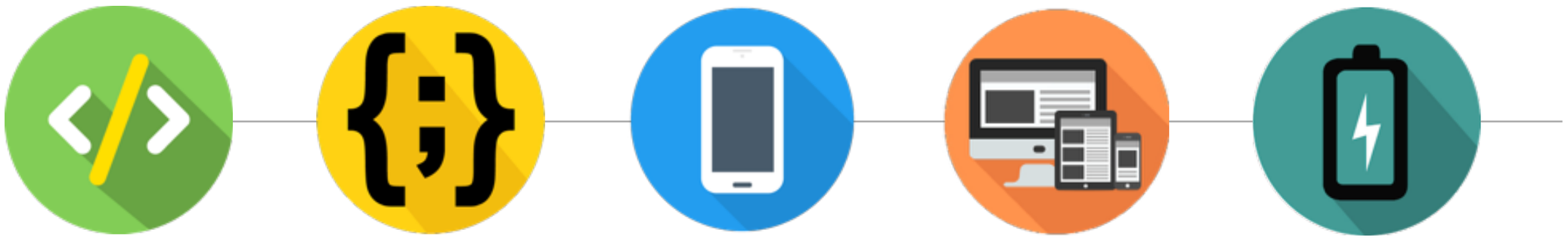
1990
HTML

1996
JavaScript

2008
Mobile Web

2012
Responsive
Web

2016
Watt Wise Web



The Web Evolution

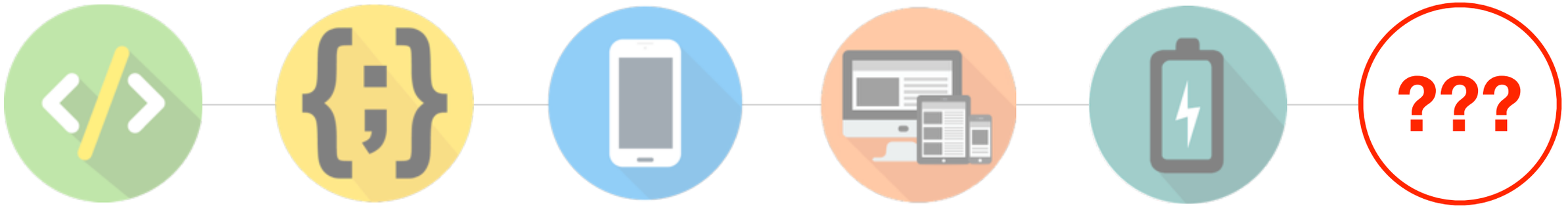
1990
HTML

1996
JavaScript

2008
Mobile Web

2012
Responsive Web

2016
Watt Wise Web



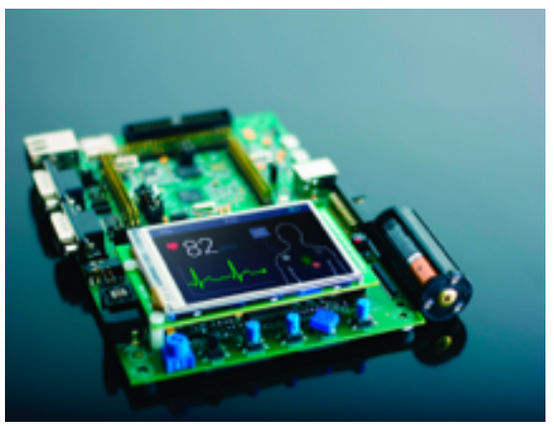

SAMSUNG
12:45
Life companion
12:45
Thu, March 14

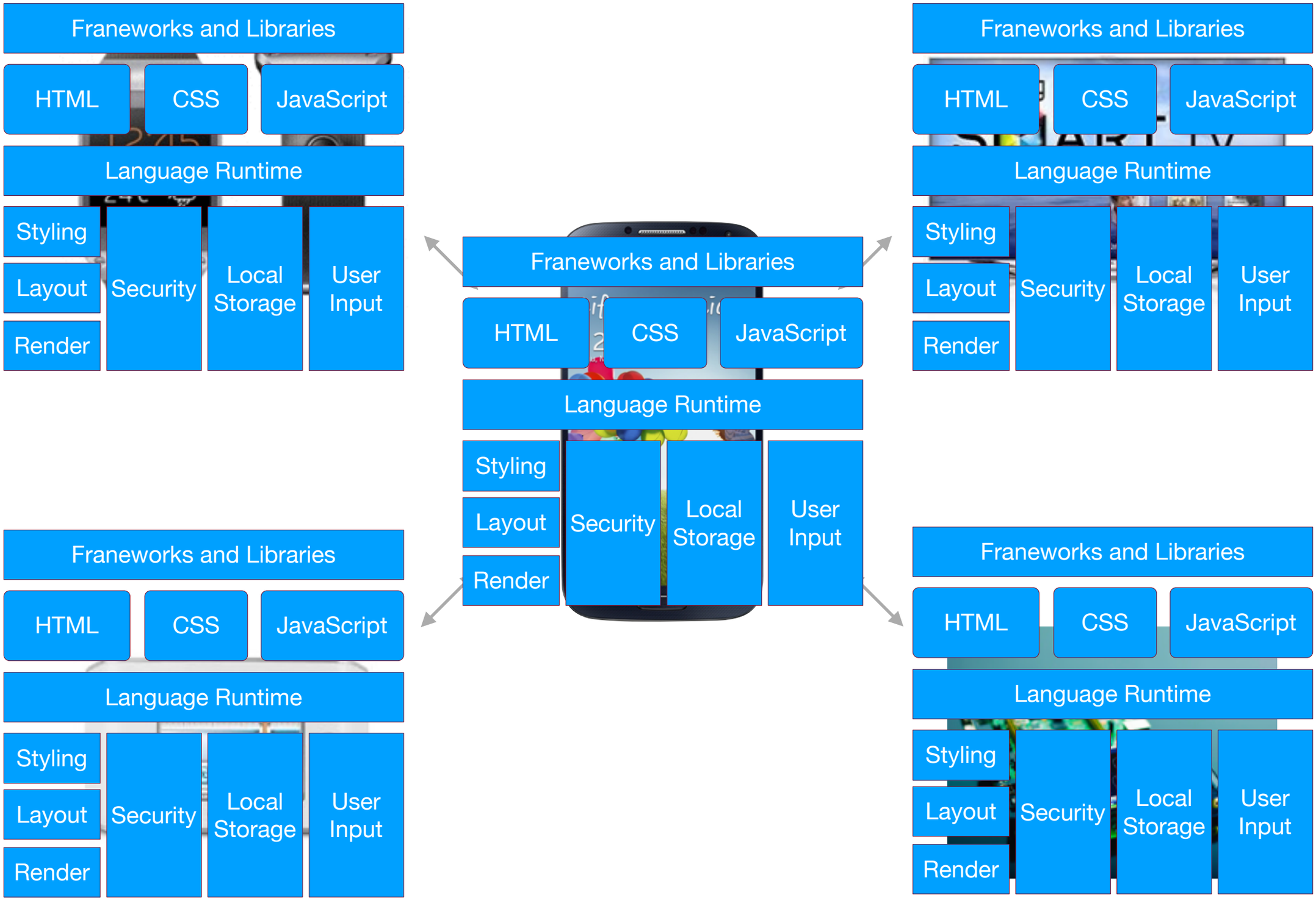

12:45
24°c
Updated 04/09 12:45


SAMSUNG
12:45
Life companion
12:45
Thu, March 14


Samsung
SMART TV
Smart Interaction
Smart Content
Smart Evolution


72
72
ecobee


82


Franeworks and Libraries
HTML
CSS
JavaScript
Language Runtime
Styling
Layout
Render
Security
Local Storage
User Input
Franeworks and Libraries
HTML
CSS
JavaScript
Language Runtime
Styling
Layout
Render
Security
Local Storage
User Input
Franeworks and Libraries
HTML
CSS
JavaScript
Language Runtime
Styling
Layout
Render
Security
Local Storage
User Input
Franeworks and Libraries
HTML
CSS
JavaScript
Language Runtime
Styling
Layout
Render
Security
Local Storage
User Input

wattwiseweb.org

Thank you!

Future Web	[**ACM Queue**] Yuhao Zhu, Vijay Janapa Reddi, "The Red future of Mobile Web Computing"
GreenWeb	[**PLDI 2016**] Yuhao Zhu, Vijay Janapa Reddi, "GreenWeb: Language Extensions for Energy-Efficient Mobile Web Computing"
WebRT	[**HPCA 2015**] Yuhao Zhu, Matthew Halpern, Vijay Janapa Reddi, "Event-Based Scheduling for Energy-Efficient QoS (eQoS) in Mobile Web Applications"
	[**HPCA 2013**] Yuhao Zhu, Vijay Janapa Reddi, "High-Performance and Energy-Efficient Mobile Web Browsing on Big/Little Systems"
	[**CAL 2012**] Yuhao Zhu, Aditya Srikanth, Jingwen Leng, Vijay Janapa Reddi, "Exploiting Webpage Characteristics for Energy-Efficient Mobile Web Browsing" (**Best of CAL**)
WebCore	[**ISCA 2014**] Yuhao Zhu, Vijay Janapa Reddi, "WebCore: Architectural Support for Mobile Web Browsing"
Motivational Studies	[**IEEE MICRO 2015**] Yuhao Zhu, Matthew Halpern, Vijay Janapa Reddi, "The Role of the CPU in Energy-Efficient Mobile Web Browsing"
	[**HPCA 2016**] Matthew Halpern, Yuhao Zhu, Vijay Janapa Reddi, "Mobile CPU's Rise to Power: Quantifying the Impact of Generational Mobile CPU Design Trends on Performance, Energy, and User Satisfaction"

Server Microarch	[**MICRO 2015**] Yuhao Zhu, Daniel Richins, Matthew Halpern, Vijay Janapa Reddi, "Microarchitectural Implications of Event-driven Server-side Web Applications" (**Top Picks Honorable Mention**)
GPGPU & IP Routing Architecture	[**DAC 2011**] Yuhao Zhu, Yangdong Deng, Yubei Chen, "Hermes: An Integrated CPU/GPU Microarchitecture for IP Routing."
	[**DAC 2010**] Bo Wang, Yuhao Zhu, Yangdong Deng, "Distributed Time, Conservative Parallel Logic Simulation on GPUs."
	[**TODAES 2011**] Yuhao Zhu, Bo Wang, Yangdong Deng, "Massively Parallel Logic Simulation with GPUs."
Tools	[**ISPASS 2015**] Matthew Halpern, Yuhao Zhu, Ramesh Peri, and Vijay Janapa Reddi, "Mosaic: Cross-platform User-interaction Record and Replay for the Fragmented Android Ecosystem."
Reliability	[**IRPS 2014**] Chen Zhou, Xiaofei Wang, Weichao Xu, Yuhao Zhu, Vijay Janapa Reddi, Chris Kim, "Estimation of Instantaneous Frequency Fluctuation in a Fast DVFS Environment Using an Empirical BTI Stress-Relaxation Model."

Coursework

Name	Instructor	Semester	SUP	Grade
COMPILERS	Keshav Pingali	Fall 2010		A
ADV EMBED MICROCONTROL SYS	Mark McDermott	Spring 2011		A-
MEMORY MANAGEMENT	Kathryn McKinley	Spring 2011	Y	A
VLSI I	Jacob Abraham	Fall 2011		A-
COMP ARCH: PARALLISM/LOCLTY	Mattan Erez	Fall 2011		A
MICROARCHITECTURE	Yale Patt	Spring 2012		B
DYNAMIC COMPILATION	Vijay Janapa Reddi	Spring 2012		A-
COMP PERF EVAL/BENCHMARKING	Lizy John	Fall 2012		B+
PARALLEL COMP ARCHITECTURE	Derek Chiou	Spring 2013		B+
HUMAN COMPUT & CROWDSRCING	Matt Lease	Fall 2015	Y	A-