

Liquid IT : Toward a better compromise between
development scalability and performance scalability
not definitive

Etienne Brodu

February 8, 2016

Abstract

TODO translate from below when ready

Résumé

Internet démultiplie nos moyens de communications. Tout en réduisant leur latence de manière à développer l'économie à l'échelle planétaire. Il permet de mettre un service à disposition de milliards d'utilisateurs en seulement quelques heures. La plupart des grands services actuels ont commencé comme de simples applications créées dans un garage par une poignée de personnes. C'est cette promesse qui a permis jusqu'à maintenant une telle croissance sur le web. Google, Facebook ou Twitter en sont quelques exemples. Au cours du développement d'une application, il est important de suivre cette croissance, au risque de se faire rattraper par la concurrence. Ce développement est guidé par les besoins en terme de fonctionnalités, afin de vérifier rapidement si le service peut satisfaire l'audience. On parle d'approche modulaire des fonctionnalités. Des langages tel que Ruby ou Java se sont imposés comme les langages du web parce qu'ils intègrent cette approche et permettent d'intégrer facilement de nouvelles fonctionnalités.

Une application qui répond correctement aux besoins atteindra de manière virale un nombre important d'utilisateurs. Son audience peut prendre plusieurs ordres de grandeurs en quelques jours voire en quelques heures si elle est correctement relayée. Une application est dite *scalable* si elle peut absorber ces augmentations d'audience. Or il est difficile pour une application d'être à la fois modulaire et *scalable*.

Au moment où l'audience devient très importante, il est souvent nécessaire de modifier l'approche de développement de l'application. Le plus souvent cela implique de la réécrire complètement en utilisant des infrastructures *scalables* qui imposent des modèles de programmation et des API spécifiques. Cela représentent une charge de travail conséquente et incertaine. De plus, l'équipe de développement doit concilier cette nouvelle approche de développement *scalable*, avec la demande en fonctionnalités. Aucun langage ne concilie ces deux objectifs. La maîtrise de ces enjeux est clé pour la pérennité de l'application.

Cette thèse est source de propositions pour écarter ce risque. Elle repose sur les deux observations suivantes. D'une part, Javascript est un langage qui a gagné en popularité ces dernières années. Il est omniprésent sur les clients, et commence à s'imposer également sur les serveurs avec Node.js. Il a accumulé une communauté de développeurs importante, et constitue l'environnement d'exécution le plus largement déployé. De ce fait, il se place maintenant de plus en plus comme le langage principal du web, détrônant Ruby ou Java. D'autre part, l'exécution de Javascript s'assimile à un pipeline. La boucle événementielle de Javascript exécute une suite de fonctions

dont l'exécution est indépendante, mais qui s'exécutent sur un seul cœur pour profiter d'une mémoire globale.

L'objectif de cette thèse est de maintenir une double représentation d'un code Javascript grâce à une équivalence entre l'approche modulaire, et l'approche pipeline d'un même programme. La première répondant aux besoins en fonctionnalités, et favorise les bonnes pratiques de développement pour une meilleure maintenabilité. La seconde propose une exécution plus efficace que la première en permettant de rendre certaines parties du code relocalisables en cours d'exécution.

Nous étudions la possibilité pour cette équivalence de transformer un code d'une approche vers l'autre. Grâce à cette transition, l'équipe de développement peut continuellement itérer le développement de l'application en suivant les deux approches à la fois, sans être cloisonné dans une, et coupé de l'autre.

Nous construisons un compilateur permettant d'identifier les fonctions de Javascript et de les isoler dans ce que nous appelons des Fluxions, contraction entre fonctions et flux. Un conteneur qui peut exécuter une fonction à la réception d'un message, et envoyer des messages pour continuer le flux vers d'autres fluxions. Les fluxions sont indépendantes, elles peuvent être déplacées d'une machine à l'autre.

Nous montrons qu'il existe une correspondance entre le programme initial, purement fonctionnel, et le programme pivot fluxionnel afin de maintenir deux versions équivalentes du code source. En ajoutant à un programme écrit en Javascript son expression en Fluxions, l'équipe de développement peut le rendre *scalable* sans effort, tout en étant capable de répondre à la demande en fonctionnalités.

Ce travail s'est fait dans le cadre d'une thèse CIFRE dans la société Worldline. L'objectif pour Worldline est de se maintenir à la pointe dans le domaine du développement et de l'hébergement logiciel à travers une activité de recherche. L'objectif pour l'équipe Dice est de conduire une activité de recherche en partenariat avec un acteur industriel.

CONTENTS

1	Introduction	8
1.1	Web development	10
1.2	Performance requirements	11
1.3	Problematic and proposal	11
1.4	Thesis organization	12
2	Context and objectives	13
2.1	The Web as a Platform	15
2.1.1	The Language of the Web	15
2.1.1.1	The Ugly Duckling	16
2.1.1.2	The Rise of Javascript	17
2.1.2	Highly Concurrent Web Servers	19
2.1.2.1	Event-driven Execution Model	20
2.1.2.2	Pipeline Execution Model	21
2.2	An Economical Problem	22
2.2.1	Disrupted Web Development	23
2.2.1.1	Power-Wall Disruption	23
2.2.1.2	Unavoidable Modularity	23
2.2.1.3	Technological Shift	24
2.2.2	Seamless Web Development	24
2.2.2.1	Real-Time Streaming Web Services	24
2.2.2.2	Differences	25
2.2.2.3	Equivalence	25

3	Software Design, State Of The Art	27
3.1	Definitions	31
3.1.1	Productivity	31
3.1.1.1	Modularity	31
3.1.1.2	Encapsulation	32
3.1.1.3	Composition	33
3.1.2	Efficiency	33
3.1.2.1	Independence	33
3.1.2.2	Atomicity	34
3.1.2.3	Granularity	34
3.1.3	Adoption	35
3.2	Productivity Focused Platforms	36
3.2.1	Modular Programming	37
3.2.1.1	Imperative Programming	37
3.2.1.2	Object Oriented Programming	37
3.2.1.3	Functional Programming	38
3.2.1.4	Multi-Paradigm	38
3.2.2	Adoption	39
3.2.2.1	Community	40
3.2.2.2	Industry	41
3.2.3	Efficiency Limitations	43
3.2.4	Summary	44
3.3	Efficiency Focused Platforms	45
3.3.1	Concurrency	45
3.3.1.1	Concurrent Programming	46
3.3.1.2	Parallel Programming	49
3.3.1.3	Summary of Concurrent and Parallel Programming Models	50
3.3.2	Adoption	51
3.3.2.1	Concurrent Programming	51
3.3.2.2	Parallel Programming	52
3.3.2.3	Stream Processing Systems	53
3.3.3	Productivity Limitations	55
3.3.4	Summary	56
3.4	Adoption Focused Platforms	57
3.4.1	Abstraction of Tasks Organization	58
3.4.1.1	Compilers	58
3.4.1.2	Runtimes	60

3.4.1.3	Productivity and Efficiency	60
3.4.2	Limitation	61
3.4.3	Summary	62
3.5	Discontinuous Developments	63
4	Seamless Shift	
	From Productivity To Efficiency	66
4.1	Proposition	68
4.1.1	Continuous Development	69
4.1.2	Equivalence	69
4.1.2.1	Execution Independence	69
4.1.2.2	Rupture points	70
4.1.2.3	Invariance	71
4.1.2.4	Transformation	72
4.2	Execution Models	74
4.2.1	Event-Driven Execution Model	74
4.2.1.1	Continuation Passing Style	75
4.2.1.2	Promise	76
4.2.2	Fluxional Execution Model	77
4.2.2.1	Fluxion	78
4.2.2.2	Streams	79
4.2.3	Examples	79
4.2.3.1	Fluxional Application	79
4.2.3.2	Transformation	81
4.3	Evaluation	83
4.3.1	Productivity and Adoption	84
4.3.2	Efficiency	84
4.3.3	Summary	85
5	Implementation	86
5.1	Dues	88
5.1.1	Definition	89
5.1.1.1	Usage	89
5.1.1.2	Creation	89
5.1.1.3	Composition	90
5.1.2	From Continuations to Dues	90
5.1.2.1	Execution order	91
5.1.2.2	Execution linearity	91

5.1.2.3	Variable scope	92
5.1.3	Due Compiler	93
5.1.3.1	Identification of continuations	93
5.1.3.2	Generation of chains	94
5.1.3.3	Evaluation	95
5.2	Fluxions	101
5.2.1	Fluxions Compiler	101
5.2.1.1	Detection	102
5.2.2	Fluxions Isolation	103
5.2.2.1	Scope	103
5.2.2.2	Stream	104
5.2.2.3	Share	104
5.2.3	Real test case	105
5.2.3.1	Compilation	106
5.2.3.2	Isolation	107
6	Conclusion	109
6.1	Summary	111
6.1.1	Fluxional Execution Model	111
6.1.2	Pipeline Extraction	112
6.1.3	Pipeline Isolation	112
6.2	Perspectives	112
6.3	Perspectives	112
6.3.1	Limitations	113
6.3.1.1	Levels of Indirections	113
6.3.1.2	Incertitudes	113
6.3.2	Dynamic Analysis and Compilation	114
6.3.2.1	Just-in-time Compilation	114
6.3.2.2	Dynamic Grouping	115
6.4	Opening	115

LIST OF FIGURES

2.1	Javascript timeline	19
2.2	Event-driven execution model	20
2.3	Pipeline execution model	22
2.4	Comparison of the two memory models	22
3.1	Balance between Efficiency and Productivity	36
3.2	Focus on Productivity	37
3.3	Steering back toward Performance Efficiency	39
3.4	TIOBE ranking	40
3.5	Most Wanted Technologies in 2015 from a StackOverflow survey	41
3.6	Languages Ranks from number of Github projects	41
3.7	StackOverflow Tags evolution	42
3.8	Module Counts per package manager	42
3.9	Focus on Efficiency	46
3.10	Steering back toward Productivity	51
3.11	Focus on Adoption	58
4.1	Differences of memory abstraction	68
4.2	Rupture point	70
4.3	Sequential scheduling	71
4.4	Causal scheduling	71
4.5	Message passing memory update	72
4.6	Sequential execution	72
4.7	Equivalence between handlers and tasks	74
4.8	Distribution of the global memory abstraction with message passing	74
4.9	Chain of continuations	75

4.10	Syntax of a high-level language to represent a program in the fluxional form	78
4.11	Screenshot from the grumpy console	80
4.12	The fluxional execution model in details	82
5.1	Roadmap	88
5.2	Simple transformation	91
5.3	Composition transformation	92
5.4	Transformation of a tree of continuations into a chain of Due	94
5.5	Results of the Due compiler evaluation	96
5.6	Compilation chain	102
5.7	Rupture point interface	103
5.8	Variable management from Javascript to the high-level fluxional language	104

LIST OF TABLES

3.1	Productivity of Modular Programming Platforms	39
3.2	Adoption of Modular Programming Platforms	43
3.3	Efficiency of Modular Programming Platforms	44
3.4	Summary of Modular Programming Platforms	45
3.5	Efficiency of Concurrent and Parallel Programming Platforms	50
3.6	Adoption of Concurrent and Parallel Programming Platforms	55
3.7	Productivity of Concurrent, Parallel and Stream Programming Platforms	56
3.8	Summary of Concurrent and Parallel Programming Platforms	57
3.9	Productivity of Compilation and Runtime Platforms	61
3.10	Efficiency of Compilation and Runtime Platforms	61
3.11	Adoption of Compilation and Runtime Platforms	62
3.12	Summary of Compilation and Runtime Platforms	63
3.13	Summary of the state of the art	64
4.1	Productivity of the proposed solution	84
4.2	Efficiency of the proposed solution	85
4.3	Sommary of the proposed solution	85

CHAPTER 1

INTRODUCTION

Contents

1.1	Web development	10
1.2	Performance requirements	11
1.3	Problematic and proposal	11
1.4	Thesis organization	12

When the 7 years old I was laid amazed eyes on the first family computer, my life goal became to know everything there is to know about computers. This thesis is a mild achievement. It compiles my PhD work on *bridging the gap between development productivity and performance efficiency, in the case of real-time web applications*.

This work is the fruit of a collaboration between the Worldline company and the Inria DICE team (Data on the Internet at the Core of the Economy) from the CITI laboratory (Centre d’Innovation en Télécommunications et Intégration de services) at INSA de Lyon. For Worldline, this work falls within a larger work named Liquid IT, on the future of the cloud infrastructure and development. As defined by Worldline, Liquid IT aims at decreasing the time to market of a web application, allows the development team to focus on application specifications rather than technical optimizations and ease maintenance. The purpose of this PhD work, was to separate development productivity from performance efficiency, to allow a continuous development from prototyping phase, until runtime on thousands of clusters. On the other hand, the DICE team focuses on the consequences of technology on economical and social changes at the digital age. This work falls within this scope as it studies the relation between the economical and the technological constraints driving the development of web applications.

1.1 Web development

Internet allows very quick releases of a minimal viable product (MVP). In a matter of hours, it is possible to release a prototype and start gathering a user community around. “*Release early, release often*”, and “*Fail fast*” are the punchlines of the web entrepreneurial community. It is crucial for the prosperity of a project to quickly validate that the proposed solution meets the needs of its users. Indeed, the lack of market need is the first reason for startup failure.¹ Often the development team quickly concretises an MVP and iterates on it using a feature-driven and monolithic approach thanks to imperative languages like Java or Ruby.

¹<https://www.cbinsights.com/blog/startup-failure-post-mortem/>

1.2 Performance requirements

If the application successfully complies with users requirements, its user base might grow with its popularity. The application is scalable when it can efficiently respond to this growth. However, it is difficult to develop scalable applications with the feature-driven approach mentioned above. Eventually this growth requires to discard the initial monolithic approach to adopt a more efficient processing model instead. Many of the most efficient models distribute the application on a cluster of commodity machines.

Once split, the application parts are connected by an asynchronous messaging system. Many tools have been developed to express and manage these parts and their communications. However, these tools impose specific interfaces and languages, different from the initial monolithic approach. It requires the development team either to be trained or to hire experts, and to start over the initial code base. This shift causes the development team to spend development resources in background without adding visible value for the users. It is a risk for the evolution of the project as the second and third reasons for startup failures are running out of cash, and missing the right competences.

1.3 Problematic and proposal

These shifts are a risk for the economical evolution of a web application by disrupting the continuity of its development process. The main question addressed by this thesis is how to avoid these shifts, so as to allow a continuous development? It implies the reconciliation between the productivity required in the early stage of development and the efficiency required with the growth of popularity. To answer this question, this thesis proposes a solution based on the equivalence between two different programming models. On one hand, there is the asynchronous, functional programming model, embodied by the Javascript event-loop. On the other hand, there is the distributed, dataflow programming model, embodied by the the pipeline architecture.

This thesis contains two main contributions. The first contribution is a compiler allowing to split a program into a pipeline of stages depending on a common memory store. The second contribution, stemming from the first one, is a second compiler,

enforcing isolation between the stages of this pipeline. With these two contributions, it is possible to transform the modular representation of an applicaiton into a pipeline representation. The modular representation allows develoment productivity, while the pipeline representation carries its execution efficiently. A development team shall then use these two representations to continuously iterate over the implementation of an application, and reach the best compromise between productivity and efficiency.

1.4 Thesis organization

This thesis is organized in six main chapters. Chapter 2 introduces the context for this thesis and explains in greater details its objectives. It presents the challenge to build web applications at a world wide scale, without jamming the organic evolution of its implementation. It concludes drawing a first answer to this challenge. Chapter 3 presents the works surrounding this thesis, and how they relate to it. It defines into the notions outlined in chapter 2 to help the reader understand better the context. The end of this chapter presents clearly the problematic addressed in this thesis. Chapter 4 introduces the proposition of this thesis, and the articulation of the contributions. Chapter 5 presents the implementations of the two contributions. The first contribution allows to represent a program as a pipeline of stages. It introduces Dues, based on Javascript Promises, to layout the pipeline. The second contribution allows to make these stages independent. It introduces Fluxions to isolate these stages, and distribute the execution. Chapter ?? evaluates this proposition at the light of the previous works, and draws the possible perspectives beyond this work. Finally, chapter 6 concludes this thesis.

CHAPTER 2

CONTEXT AND OBJECTIVES

Contents

2.1	The Web as a Platform	15
2.1.1	The Language of the Web	15
2.1.1.1	The Ugly Duckling	16
2.1.1.2	The Rise of Javascript	17
2.1.2	Highly Concurrent Web Servers	19
2.1.2.1	Event-driven Execution Model	20
2.1.2.2	Pipeline Execution Model	21
2.2	An Economical Problem	22
2.2.1	Disrupted Web Development	23
2.2.1.1	Power-Wall Disruption	23
2.2.1.2	Unavoidable Modularity	23
2.2.1.3	Technological Shift	24
2.2.2	Seamless Web Development	24
2.2.2.1	Real-Time Streaming Web Services	24
2.2.2.2	Differences	25
2.2.2.3	Equivalence	25

The web allows applications to grow a user base very quickly. The development of a web application needs to follow this rapid pace to assure performance efficiency. But the languages often fail to grow with the project they initially supported very efficiently.

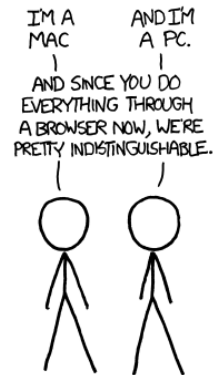
The inadequacy of the languages to support the growth of web applications leads to wasted development efforts, and additional costs. The objective of this thesis is to avoid these efforts and costs. It intends to provide a continuous development from the initial prototype up to the releasing and maintenance of the complete product.

This chapter presents the general context for this work, and defines the scope of this thesis. Section 2.1 presents the motivations that led the web to become a software platform, and the context of web development. It presents Javascript, one of the most important language in web development. Then, it presents the challenges of developing web servers for large audiences. Section 2.2 states the problem tackled by this thesis, and its objectives.

2.1 The Web as a Platform

Similarly to operating systems, Web browsers started as software products with extension capabilities that transformed it into a platform. The distribution of an application is limited only by the platform it can be deployed on. The Web spreads the scalability of software distribution world wide with a near zero latency. It eventually became the main distribution medium, and the wider market there can possibly be for software. It led the Web to become a major platform, replacing operating systems.

Now, with web applications, the distribution medium is so transparent that owning a software product to have an easier access is no longer relevant. It stimulates a completely new business model based on an instantaneous and free access for the user, while claiming value for their data.



2.1.1 The Language of the Web

In the 80's, reducing development time became more profitable than reducing hardware costs. Higher-level languages replaced lower-level languages. The economical

gain in development time brought by productive languages compensated the decrease in performance. Most of the now popular programming languages were released at this time, Python in 1991, Ruby in 1993, Java in 1994, PHP and Javascript in 1995.

With the democratisation of programming, the involvement of a community became critical for the adoption, evolution and maturation of a language. Communities adopts a language because it allows to quickly experiment and enter businesses sectors. The industry adopts a language because it responds to business needs and its community represents a hiring pool. The community support and the industrial needs are reinforcing each other in a loop.

Java thrived in the software industry, but lose the hype that drove the community innovation and creativity. Now, it struggles to keep up with the latest trends in software development. On the contrary, Ruby on Rails emerged from an industrial context, but is now open source, and backed by a strong community that makes it evolve and mature. Other languages like Python and PHP, emerged within a strong community, and were later adopted by the industry for web development. Django, the Python web frameworks, is used to develop many web applications in industrial contexts. The Wordpress publishing platform is another example of an economical success with PHP.

The web acts as a catalyst in the interaction between the community and the industry. Because of its position in the web, Javascript is slowly becoming the main language for web development. The next paragraph present its evolution in the industry and the community.

2.1.1.1 The Ugly Duckling

“*There are only two kinds of languages: the ones people complain about and the ones nobody uses*”

— B. Stroustrup¹

Javascript was released as a scripting engine in Netscape Navigator around September 1995 and later in its concurrent, Internet Explorer. The differences between the

¹http://www.stroustrup.com/bs_faq.html#really-say-that

two implementations forced Web pages to be designed for a specific browser. This competition was fragmenting the Web. To stop this fragmentation, Netscape submitted Javascript to ECMA International for standardization in November 1996. ECMA International released ECMAScript – or ECMA-262 – the first standard for Javascript in June 1997.

The initial release of Javascript was designed by Brendan Eich within 10 days, and targeted unexperienced developers. For these reasons, the language was considered poorly designed and unattractive by the developer community.

But this situation evolved drastically since. All web browsers include a Javascript interpreter, making Javascript the most ubiquitous runtime [51]. This position became an incentive to make it fast (V8, ASM.js) and convenient (ES6, ES7). Any Javascript code in the browser is open, allowing the community to pick, improve and reproduce the best techniques². Javascript is distributed freely, with all the tools needed to reproduce and experiment on the largest communication network in history. And since 2009, it came back on the server³ with Node.js. This omnipresence became an advantage. It allows to develop and maintain the whole application with the same language. All these reasons made the popularity of the Web and Javascript.



2.1.1.2 The Rise of Javascript

“*When JavaScript was first introduced, I dismissed it as being not worth my attention. Much later, I took another look at it and discovered that hidden in the browser was an excellent programming language.*”

²<http://blog.codinghorror.com/the-power-of-view-source/>

³True hipsters used Server-Side Javascript before it was cool. https://developer.mozilla.org/en-US/docs/Archive/Web/Server-Side_JavaScript

Javascript was initially used for short interactions on web pages. Nowadays, there are a lot of web-based applications replacing desktop applications, like mail client, word processor, music player, graphics editor...

ECMA International allowed this progression by releasing several versions to give Javascript a more complete and solid base as a programming language. Moreover, Asynchronous Javascript And XML (Ajax) allows to dynamically reload the content inside a web page, hence improving the user experience [59]. It allows Javascript to develop richer applications inside the browser, from user interactions to network communications. The community released frameworks to assist the development of these larger applications. Prototype⁵ and DOJO⁶ are early famous examples, and later jQuery⁷ and underscore⁸.

Since 2004, the Web Hypertext Application Technology Working Group⁹ worked on the fifth version of the HTML standard. The name is misleading, it is really about giving Javascript superpowers like geolocation, storage, audio, video, and many mores. The simultaneous release of HTML5, ECMAScript 5 and V8, around 2009, represent a milestone in the development of web-based applications. Javascript became the *de facto* programming language to develop on this rising application platform that is the Web¹⁰. The milestones in the history of Javascript presented in the previous chapters are summarized in figure ??.

×

Javascript is now widely used on the web, in open source projects, and in the software industry. With the increasing importance of client web applications, Javascript

⁴<http://javascript.crockford.com/survey.html>

⁵<http://prototypejs.org/>

⁶<https://dojotoolkit.org/>

⁷<https://jquery.com/>

⁸<http://underscorejs.org/>

⁹<https://whatwg.org/>

¹⁰<http://blog.codinghorror.com/javascript-the-lingua-franca-of-the-web/>

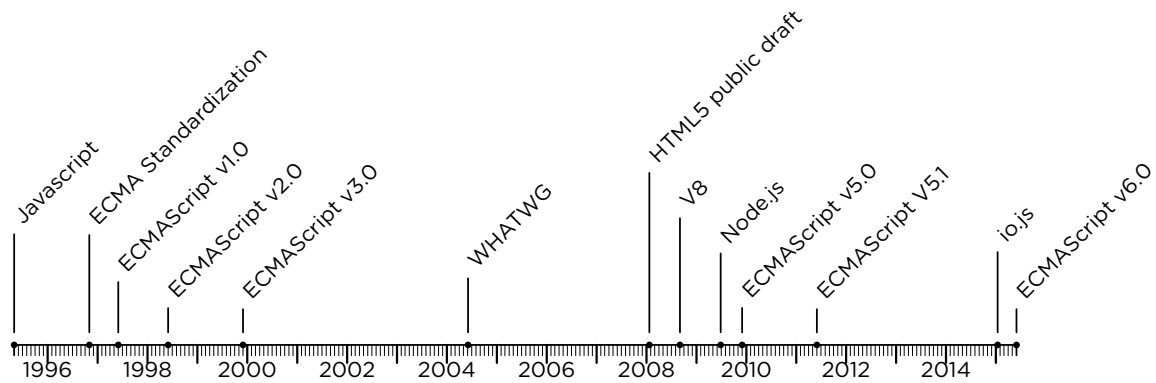


Figure 2.1 – Javascript timeline

is assuredly one of the most important language in the times to come. Especially that Javascript now allows to build the server side of web applications as well. The next section presents the realities and technical challenges to assure the performance of web applications against billions of users.

2.1.2 Highly Concurrent Web Servers

Since the web allowed an application to scale world wide with near zero latency, the software industry needed innovative solutions to cope with large network traffic.

The Internet allows communication at an unprecedented scale. There is more than 16 billions connected devices, and it is growing fast¹¹ [87]. A large web application like google search receives about 40,000 requests per seconds¹². Such a Web application needs to be highly concurrent to manage this amount of simultaneous requests. In the 2000s, the limit to break was 10 thousands simultaneous connections with a single commodity machine¹³. In the 2010s, the limit is set at 10 millions simultaneous connections¹⁴. With the growing number of connected devices on the internet, concurrency is a very important property in the design of web applications.

2.1.2.1 Event-driven Execution Model

¹¹<http://blogs.cisco.com/news/cisco-connections-counter>

¹²<http://www.internetlivestats.com/google-search-statistics/>

¹³<http://www.kegel.com/c10k.html>

¹⁴<http://c10m.robertgraham.com/p/manifesto.html>

Javascript is often associated with an event-driven paradigm to react to concurrent user interactions. This paradigm proved to be very efficient as well for a web application to react to concurrent requests. In 2009, Joyent released Node.js to build real-time web applications with this paradigm.

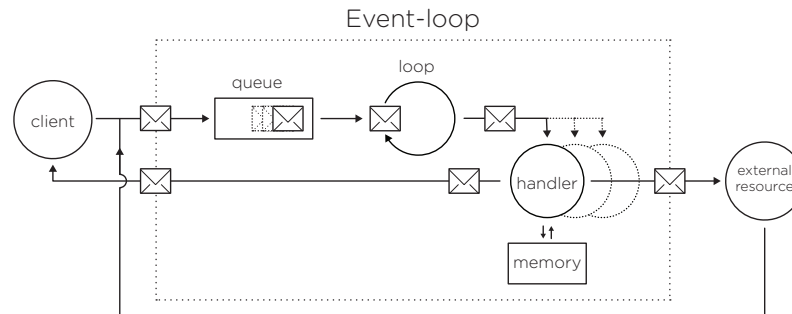


Figure 2.2 - Event-driven execution model

The event-driven execution model is presented in figure ?? . At reception, each request from a client queues an event waiting to be processed. A loop unqueues these events one at a time, and run the appropriate handler to process them. To process an event, a handler can query remote resources, which respond asynchronously by queuing additional events. The querying handler specifies new handlers - called continuations - to process these additional events. Alternatively, a handler can respond directly to the client, ending this chain of asynchronous events.

This execution model allows the high concurrency required to respond to a high number of users simultaneously. This concurrency needs to be scalable to adapt to the growth of audience, as explained in the next paragraph.

Scalability

The traffic of a popular web application such as Google search remains stable, while the traffic of a less popular web application is much more uncertain. Moreover, the load of the web application grows with its user base. The available resources need to increase as well to meet this load. For stable traffic, this growth is steady enough to plan the increase of resources ahead of time. But for unstable traffic, it is erratic and challenging to meet.

An application is scalable, if it is able to spread over resources proportionally as a reaction to its load to use these resources efficiently. It is a desirable property, as it

helps to meet the growth, without spending time to manually spread the application on available resources to react to this erratic growth.

Time-slicing and Parallelism

Concurrency is achieved differently on hardware with a single or several processing units. On a single processing unit, the tasks are executed sequentially, interleaved in time. While on several processing units, the tasks are executed simultaneously, in parallel. Parallel executions uses more processing units to reduce computing time over sequential execution.

If the tasks are independent, they can be executed in parallel as well as sequentially. This parallelism is scalable, as the independent tasks can stretch the computation on the resources so as to meet the required performance. However, the tasks within an application need to coordinate together to modify the application state. This coordination limits the parallelism and imposes to execute some tasks sequentially. It limits the scalability. The type of possible concurrency, sequential or parallel, is defined by the interdependencies of the tasks.

The Javascript event-loop requires a global memory to assure the interdependency of the tasks. This thesis argues that there exists an equivalence between the event-driven model and the pipeline execution model.

2.1.2.2 Pipeline Execution Model

The pipeline execution model, presented in figure ??, is composed of isolated stages communicating by message passing to leverage the parallelism of a multi-core hardware architectures. It is well suited for streaming application, as the stream of data flows from stage to stage. Each stage has an independent memory to hold its own state. As the stages are independent, the state coordination between the stages are communicated along with the stream of data.

The execution model of each stage is organized in a similar fashion than the event-loop presented previously. It receives and queues messages from upstream stages, processes them one after the other, and outputs the result to downstream stages. The pipeline architecture is different as each task is executed on an isolated stage. Whereas in the event-driven execution model, all handlers share the same queue, loop and memory store. This difference is illustrated in figure ?. The isolation of memory in the pipeline execution model impacts the productivity of its programming model.

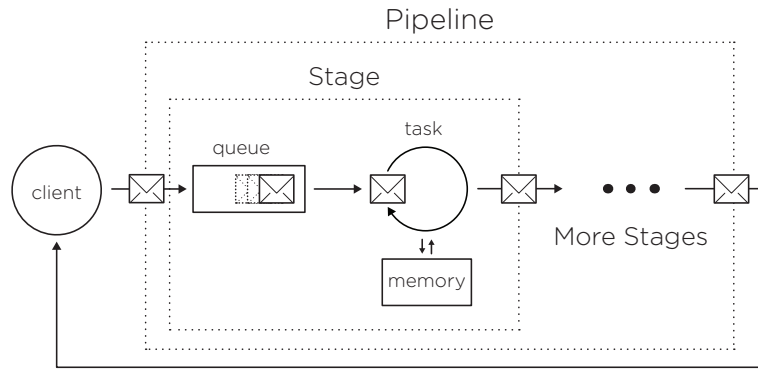


Figure 2.3 – Pipeline execution model

The next section details further the incompatibility between the two programming model and the resulting economical consequences.

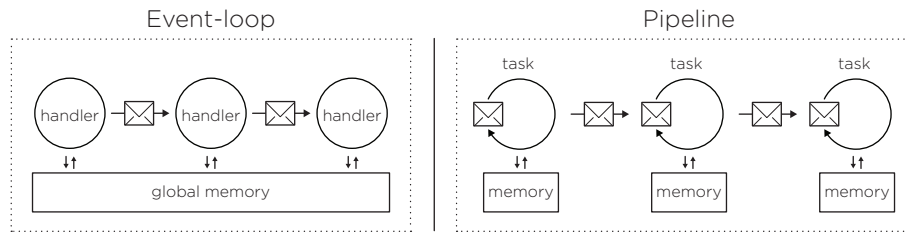


Figure 2.4 – Comparison of the two memory models

2.2 An Economical Problem

With SaaS, the software industry is in charge of both development and execution of the software. The previous section presented these two aspects individually. This section presents the challenges encountered by conducting the two at world wide scale. It then focuses on the subject and defines the objectives of this thesis.

2.2.1 Disrupted Web Development

The economical constraints to meet are very different in the beginning and during the maturation of a web application. In the early steps the constraints hold on the development productivity. The team needs to reduce development costs, and to release a first version as soon as possible. On the contrary, during the maturation of the application, the constraints hold on the performance efficiency. The application needs to be highly concurrent to meet the load of usage.

The team need to revise its approach to meet these different constraints. These revisions leads to disruptions in the evolution of the application.

2.2.1.1 Power-Wall Disruption

Around 2004, manufacturers reached what they called the *Power-wall*. The speed of sequential execution on a processing unit plateaued¹⁵. Therefore, the performance of sequential programming plateaued as well. They started to arrange transistors into several processing units to keep increasing overall performance efficiency. Parallel programming became the only option to achieve high concurrency, but the memory isolation it requires limits the productivity. This *Power-wall* leads to a rupture between efficiency and productivity.

2.2.1.2 Unavoidable Modularity

The best practices for productivity in software development advocate to gather features logically into distinct modules. This modularity allows a developer to understand and contribute to an application one module at a time, instead of understanding the whole application. It allows to develop and maintain a large code-base by a multitude of developers bringing small, independent contributions.

This modularity avoids a different problem than the isolation required by parallelism. The former intends to structure code to improve maintainability, while the latter improve performance through parallel execution. These two organizations are conflicting in the design of the application. The next paragraph presents the disruptions in the development of a web application implied by this conflict.

¹⁵<https://cartesianproduct.wordpress.com/2013/04/15/the-end-of-dennard-scaling/>

2.2.1.3 Technological Shift

The development team opt for a popular and accessible language to be productive in the beginning of the project. It is only after a certain threshold of user load that the economical constraint on efficiency exceeds the one on productivity. The development team then shifts to an organization providing parallelism.

This shift brings two risks. The development team needs to rewrite the code base to adapt it to a completely different paradigm. The application risks to fail because of this challenge. And after this shift, the development pace slows down. The development team cannot react as quickly to user feedbacks to adapt the application to the market needs. The application risks to fall in obsolescence.

The risks implied by this rupture proves that there is economically a need for a solution that continuously follows the evolution of a web application. The proposition of this thesis is presented in the next section. The proposed solution would allow developers to iterate continuously on the implementation focusing simultaneously on performance, and on maintainability.

2.2.2 Seamless Web Development _____

This thesis is conducted in the frame of a larger work on LiquidIT within the Worldline company. Worldline develops and hosts real-time streaming Web services, and identified that one of their need was to increase the time to market for its products. Worldline defines LiquidIT as *a concept of flexible and cost-effective IT services that can be provisioned, built and configured in real time, allowing end-to-end financial transparency*. It precisely intends to provide *business agility, investment-free charging models, flexibility and ease of use*. This thesis intends to allow the developer to focus solely on business logic, and leave the technical constraints of performance scalability to automated tools. The objective of this work are avoid the disruption in development, and provide a seamless development experience. They are presented in the next paragraphs.

2.2.2.1 Real-Time Streaming Web Services

This thesis focuses on web applications processing streams of requests from users in soft real-time. Such applications receive requests from clients through the HTTP

protocol and must respond within a finite window of time. They are generally organized as sequences of tasks to modify the input stream of requests to produce the output stream of responses. The stream of requests flows through the tasks, and is not stored. On the other hand, the state of the application remains in memory to impact the future behaviors of the application. This state might be shared by several tasks within the application, and imply coordination between them.

As presented in the previous section, such applications are often implemented with the event-driven programming model or the pipeline programming model. Despite the differences between the two models, an equivalence to map these differences is developed throughout this thesis.

2.2.2.2 Differences

Both programming models encapsulate the execution in tasks assured to have an exclusive access to the memory. However, they use two different models to provide this exclusivity. Contrary to the pipeline architecture, the event-loop provides a common memory store allowing the best practice of software development to improve maintainability.

However, these two organizations are incompatible. Because of economical constraints, this incompatibility implies ruptures in the development. It represents additional development efforts and important costs. This thesis argues that it is possible to allow a continuous development between the two organizations, so as to lift these efforts and costs. The argumentation of this possibility is based on an equivalence bridging the two organizations. This equivalence is presented briefly in the next paragraph, and detailed further in the chapter 4 and 5.

2.2.2.3 Equivalence

In the beginning of a project, the team adopts the event-driven execution model to focus on maintainability and evolution, discarding the scalable performance concerns. And as the project gather audience and the performance concerns become more and more critical, the development team adopt the pipeline execution model to take into account this performance concerns. The equivalence would allow a compiler to transform an application expressed in one model into the other.

With this equivalence, it would be possible to express an application following the design principles of software development. A development team could rely on

the common memory store of the event-driven execution model, and focuses on the maintainability of the implementation. And yet, because of the equivalence between these two models, the execution engine could adapt itself to any parallelism of the computing machine, from a single core, to a distributed cluster. The development team could continuously progress with the two models and take advantage of their different concerns about the implementation, performance and maintainability.

×

This thesis proposes to provide an equivalence between the two memory models for streaming web applications. The goal of conciliating these two concerns is not new. The next chapter presents all the previous results needed to understand this work, up to the latest advances in the field.

CHAPTER 3

SOFTWARE DESIGN, STATE OF THE ART

Contents

3.1	Definitions	31
3.1.1	Productivity	31
3.1.1.1	Modularity	31
3.1.1.2	Encapsulation	32
3.1.1.3	Composition	33
3.1.2	Efficiency	33
3.1.2.1	Independence	33
3.1.2.2	Atomicity	34
3.1.2.3	Granularity	34
3.1.3	Adoption	35
3.2	Productivity Focused Platforms	36
3.2.1	Modular Programming	37
3.2.1.1	Imperative Programming	37
3.2.1.2	Object Oriented Programming	37
3.2.1.3	Functional Programming	38
3.2.1.4	Multi-Paradigm	38
3.2.2	Adoption	39
3.2.2.1	Community	40
3.2.2.2	Industry	41
3.2.3	Efficiency Limitations	43
3.2.4	Summary	44
3.3	Efficiency Focused Platforms	45
3.3.1	Concurrency	45
3.3.1.1	Concurrent Programming	46

3.3.1.2	Parallel Programming	49
3.3.1.3	Summary of Concurrent and Parallel Program- ming Models	50
3.3.2	Adoption	51
3.3.2.1	Concurrent Programming	51
3.3.2.2	Parallel Programming	52
3.3.2.3	Stream Processing Systems	53
3.3.3	Productivity Limitations	55
3.3.4	Summary	56
3.4	Adoption Focused Platforms	57
3.4.1	Abstraction of Tasks Organization	58
3.4.1.1	Compilers	58
3.4.1.2	Runtimes	60
3.4.1.3	Productivity and Efficiency	60
3.4.2	Limitation	61
3.4.3	Summary	62
3.5	Discontinuous Developments	63

“
A designer is responsible for producing the greatest benefit for any given investment of time, talent, money, and other resources.”

— K. Sullivan, W. Griswold, Y. Cai, B. Hallen [151]

With the growth of Software as a Service (SaaS) on the web, the same company carries both development and exploitation of an application at scale of unprecedented size. It revealed the importance of previously unknown economic constraints. To assure the continuous growth and sustainability of an application, it needs to address two contradictory goals : development productivity and performance efficiency. These goals needs to be enforced by the platform supporting the application to build good development habits for the developers. A platform designates any solution that allows to build an application on top of it, including programming languages, compilers, interpreters, frameworks, runtime libraries and so on.

The productivity of a platform is the degree to which developers can quickly produce new and modify existing software. It impacts the maintainability of the applications and relies on the modularity enforced by its platform. *75% of your budget is dedicated to software maintenance.*¹ Especially, higher order programming is crucial to build and compose modules productively. It relies either on mutable states, or immutable states, but hardly on a combination of both.

However, neither mutable nor immutable states allows performance efficiency. Mutable states leads to synchronization overhead at a coarser-grain level, while immutable states leads to communication overhead at a finer-grain level. Efficiency relies on a combination of synchronization at a fine-grain level, and immutable message passing at a coarse-grain level. This combination breaks the modularity, hence the productivity of an application. A company has no choice but to commit huge development efforts to get efficient performances.

Moreover, a balance between productivity and efficiency is required for a platform to enter a virtuous circle of adoption. The productivity is required to be appealing to gather a community to support the ecosystem around the platform. This community is appealing for the industry as a hiring pool. Additionally, the efficiency is required

¹<http://www.castsoftware.com/glossary/software-maintainability>

to be adopted by the industry to be economically viable. And the industrial relevance provides the reason for this ecosystem to exist and the community to gather.

This chapter presents a broad view of the state of the art in the compromises between productivity and efficiency. It defines software productivity, efficiency, and adoption in section 3.1 and all the underlying concepts, such as higher order programming and state mutability. It then analyzes different platforms according to their focus. platforms focusing on productivity are addressed in section 3.2, those focusing on efficiency in section 3.3 and those focusing on a compromise between the two in section 3.4.

3.1 Definitions

The continuous growth and sustainability of a platform relies on three criteria, Productivity, Efficiency and Adoption. This section defines these tree criteria and their underlying concepts.

3.1.1 Productivity

The productivity of a platform is the degree to which developers can quickly produce new and modify existing software. To support productivity, a platform needs to enforce modularity directly in the design of applications. Productivity leads to maintainability.

Still, measuring productivity precisely to infer the improvement brought by a platform is highly difficult. The measurement varies more between developers than between platforms [141].

3.1.1.1 Modularity

Modularity is about encapsulating subproblems and composing them to allow greater design to emerge. It allows to limit the understanding required to contribute to a module [149]. And it reduces development time by allowing developers to implement different modules simultaneously [172, 22].

The criteria to define modules to improve productivity are high cohesion enforced by encapsulation and low coupling enforced by composition [149]. Cohesion defines how strongly the features inside a module are related. Coupling defines the strength of the interdependences between modules.

×

The criteria for productivity are the encapsulation and composition allowed by a platform. Encapsulation relies on the definition of boundaries, and the protection of data. Composition relies on higher-order programming and lazy evaluation. The next paragraphs define these requirements.

- Encapsulation (Boundary definition, Data protection)
→ increases Cohesion
- Composition (Higher-order programming, Lambda Expressions)
→ decreases Coupling

3.1.1.2 Encapsulation

Boundary Definition

Modular Programming draws clear interfaces around a piece of implementation so that the execution remains enclosed inside [42]. At a fine level, it helps avoid spaghetti code [45], and at a coarser level, it structures the implementation [46] into modules, or layers.

Data Protection

Modular programming encapsulates a specific design choice in each module, so that it is responsible for one and only one concern. It isolates its evolution from impacting the rest of the implementation [130, 156, 94]. Examples of such separation of concerns are the separation of the form and the content in HTML / CSS, or the OSI model for the network stack.

3.1.1.3 Composition

Higher-Order Programming

Higher-order programming introduces lambda expressions, functions manipulable like any other primary value. They can be stored in variables, or be passed as arguments. It replaces the need for most modern object oriented programming design patterns ² with Inversion of Control [98], the Hollywood Principle [154], or Monads [167]. Higher-order programming help loosen coupling, thus improve productivity [78].

Closures

In languages allowing mutable state, lambda expressions are implemented with closure to preserve the lexical scope [153]. A closure is the association of a function and a reference to the lexical context from its creation. It allows this function to access variable from this context, even when invoked outside the scope of this context.

3.1.2 Efficiency _____

The efficiency of a software project is the relation between the usage made of available resources and the delivered performances. For an application to perform efficiently, its platform needs to enforce scalability directly in its design.

Scalability relies on the parallelism allowed by the commutativity of operations execution [32]. An operation is a sequence of statements. Operations are commutative if the order of their executions is irrelevant for the correctness of their results. Commutativity assures the independence of operations.

3.1.2.1 Independence

The independence, and commutativity of an operation depends on its accesses to shared state. If the operation doesn't rely on any shared state, it is independent.

²<http://stackoverflow.com/a/5797892/933670>

The independence of operations allows to execute them in parallel, hence to increase performance proportionally to occupied resources [7, 69]. But if they rely on shared state, they need to coordinate the causal scheduling and atomicity of their executions to avoid conflicting accesses. This scheduling between the operations can be defined in two ways.

Synchronization Operations are scheduled sequentially to have the exclusivity on a shared state, or

Message-passing Operations communicate their local modifications of the state to other operations as immutable messaged.

Because of the latency associated with message-passing, the atomicity of operations is challenged.

3.1.2.2 Atomicity

An operation is atomic if it happens in a single bulk. The beginning and end are indistinguishable for an external observer. It assures the developer of the invariance of the memory during the operation. It relies either on the causal scheduling of operations – synchronization – or exclusivity of their memory accesses – message-passing.

3.1.2.3 Granularity

If the operations access the state too frequently, the communication overhead of message passing exceeds the performance gains of parallelism. Whereas if operations access the state too rarely, the synchronization required for sharing state limits the possible parallelism. These two extremes are inefficient. Operations tend to share state more closely at a fine-grain level and more loosely at a coarser-grain level. Therefore, efficiency requires the combination of fine-level state sharing to avoid communication overhead, and coarse-level independence to allow parallelization [71, 70, 123, 68]. The threshold determining frequent or rare access to the state determines the granularity level between synchronization and parallelization of tasks.

×

The criteria to analyze the performance efficiency of platforms are the synchronization available at a fine-level, and the message-passing available at a coarse-level.

- Fine-level Synchronization
→ avoids communication overhead
- Coarse-level Message-passing
→ allows parallelism

3.1.3 Adoption

An application is sustainable only if the platform used to build it generates reinforcing interactions between a community of passionate and the industry. A platform needs to present a balance between productivity and efficiency to be adopted by both the community and the industry. The productivity is required for a platform to be appealing to gather a community to support the ecosystem around it. And the efficiency is required to be economically viable and needed by the industry, and to provide the reason for this ecosystem to exist. Additionally, the web acts as a tremendous catalyst fueling these interactions.

×

The criteria to analyze the adoption of platforms are the support of the community, and the industrial need.

- Community Support
→ grows an ecosystem
- Industrial Need
→ gives a goal for this ecosystem to grow

×

Adoption requires a balance between efficiency and productivity. This incentive to balance between productivity and efficiency is illustrated in figure 3.1. This figure is used throughout this chapter to graphically represent all the platforms analyzed.

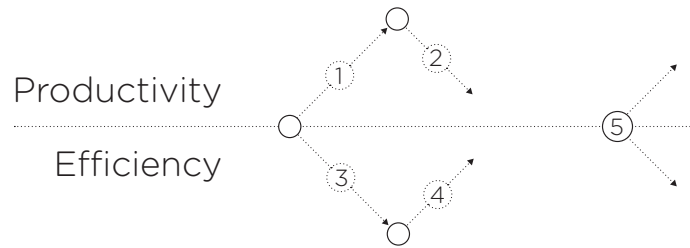


Figure 3.1 – Balance between Efficiency and Productivity

3.2 Productivity Focused Platforms

“*It is becoming increasingly important to the data-processing industry to be able to produce [programming systems] at a faster rate, and in a way that modifications can be accomplished easily and quickly.*”

— W. Stevens, G. Myers, L. Constantine [149].

In order to improve and maintain a software system, it is important to hold in mind a mental representation of its implementation [145]. As the system grows in size, the mental representation becomes more and more difficult to grasp. Therefore, it is crucial to decompose the system into smaller subsystems easier to grasp individually.

“*Measuring programming progress by lines of code is like measuring aircraft building progress by weight.*”

— Bill Gates

Section 3.2.1 presents the modular programming paradigms, and their programming models, oriented toward productivity. Section 3.2.2 presents the adoption of the implementations of modular programming languages. Section ?? presents the consequences of the modularity on performance. Finally, section 3.2.4 summarizes the three previous sections in a table.

3.2.1 Modular Programming ---

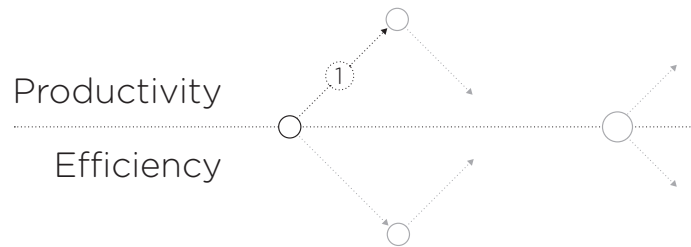


Figure 3.2 – Focus on Productivity

The next paragraphs presents the different programming model regarding their support to modular programming and productivity.

3.2.1.1 Imperative Programming

Imperative programming is the very first programming paradigm, as it evolves directly from the hardware architectures. It allows to express the suite of operation to carry sequentially on the computing processor. Most imperative languages provide encapsulation with modules but not higher-order programming. The implementations of Imperative Programming are Fortran, Algol, Cobol and C.

3.2.1.2 Object Oriented Programming

The very first Object-Oriented Programming (OOP) language was Smalltalk [61]. It defined the core concepts as message passing and encapsulation ³. Nowadays, the

³http://userpage.fu-berlin.de/~ram/pub/pub_jf47ht81Ht/doc_kay_oop_en

emblematic figures in the software industry are C++ [150], Java [63] and Pascal. They provide encapsulation with Classes, and allows mutable structures for performance reasons. They recently introduced higher-order programming with lambda expressions.

3.2.1.3 Functional Programming

The definition of pure Functional Programming resides in manipulating only expressions and replacing state mutability, with immutable message-passing. The absence of state mutability makes a function side-effect free, hence their execution can be scheduled in parallel. The most important pure Functional Programming languages are Scheme [139], Miranda [162], Haskell [92] and Standard ML [119]. They provide encapsulation, higher-order programming and lazy evaluation.

3.2.1.4 Multi-Paradigm

The functional programming concepts are also implemented in other languages along with mutable states and object-oriented concepts. Major recent programming languages, including Java 8 and C++ 11, now commonly present **higher-order functions**. *In fine*, it helps developers to write applications that are more maintainable, and favorable to evolution [93, 163]. These recent multi-paradigms languages such as Javascript, Python and Ruby combine the different paradigms to help developer building applications faster.

×

The previously presented programming models are all rather focused on productivity, as recapped in table 3.1.

Model	Composition	Encapsulation	→	Productivity
Imperative Programming	3	4		3
Object-Oriented Programming	5	5		5
Functional Programming	5	5		5
Multi Paradigm	5	5		5

Table 3.1 – Productivity of Modular Programming Platforms

3.2.2 Adoption

As stated previously, adoption relies on productivity as well as efficiency. The next paragraphs present the adoptions of different platforms for web development and focus on Javascript.

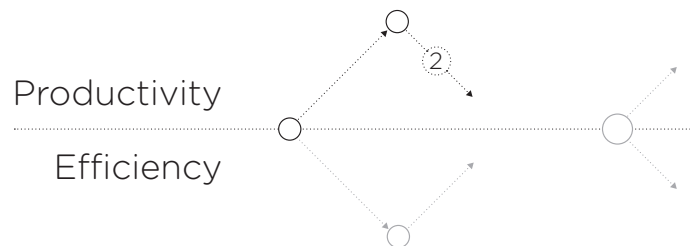


Figure 3.3 – Steering back toward Performance Efficiency

3.2.2.1 Community

Available Resources

As of December 2015, Javascript ranks 8th according to the TIOBE Programming Community index, and was the most rising language in 2014. This index measure the popularity of a programming language with the number of results on many search engines. And it ranks 7th on the PYPL. The PYPL index is based on Google trends to measure the number of requests on a programming language.

From these indexes, the major programming languages are Java, C++, C, C# and Python. These languages are still widely used by their communities and in the industry.

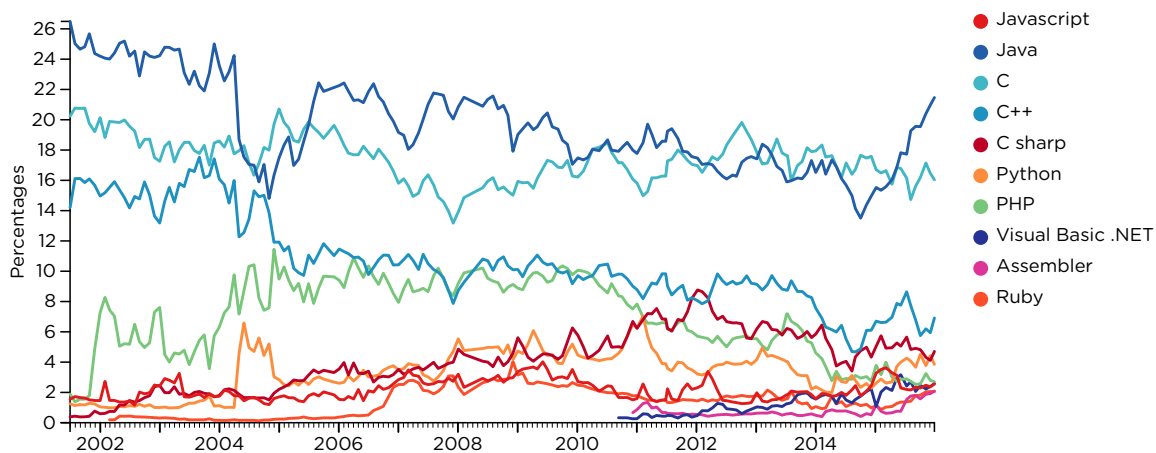


Figure 3.4 – TIOBE ranking

Developers Collaboration Platforms

Online collaboration tools give an indicator of the number of developers and projects using certain languages. Javascript is the most used language on *Github*⁴ and the most cited language on *StackOverflow*⁵. It represents more than 320,000 repositories on *Github*. The second language is Java with more than 220,000 repositories. It is cited in more than 960,000 questions on *StackOverflow* while the second is Java with around 940,000 questions. And according to a survey by *StackOverflow*,

⁴themostimportantcollaborativedevelopmentplatformgatheringabout9millionsusers.

⁵themostimportantQ&Aplatformfordevelopers.

it is currently the language the most popular⁶. Moreover, the Javascript package manager, *npm*, has the most important and impressive package repository growth.

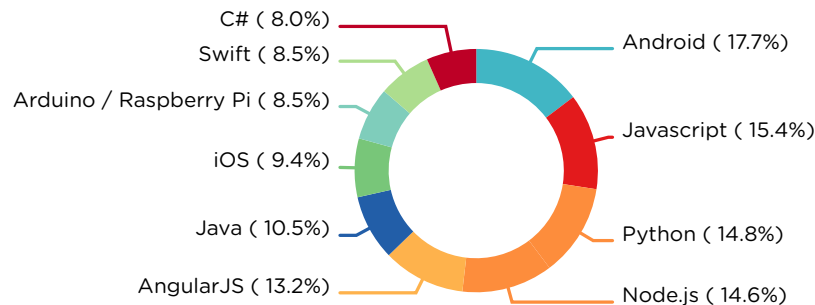


Figure 3.5 – Most Wanted Technologies in 2015 from a StackOverflow survey

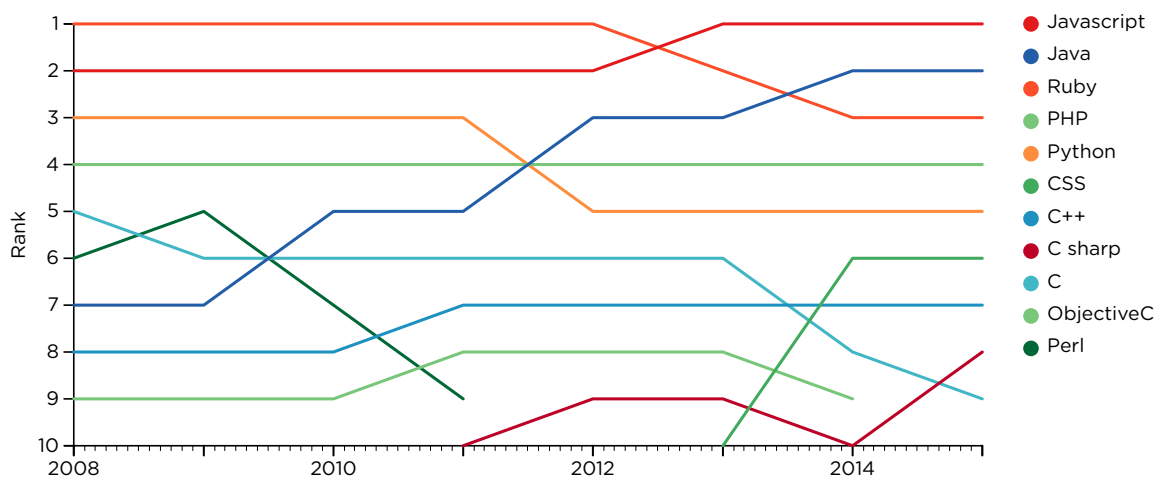


Figure 3.6 – Languages Ranks from number of Github projects

3.2.2.2 Industry

The actors of the software industry tends to hide their activities trying to keep an edge on the competition. The previous metrics represent the visible activity but are barely representative of the software industry. The trends on job opportunities give

⁶<http://stackoverflow.com/research/developer-survey-2015>

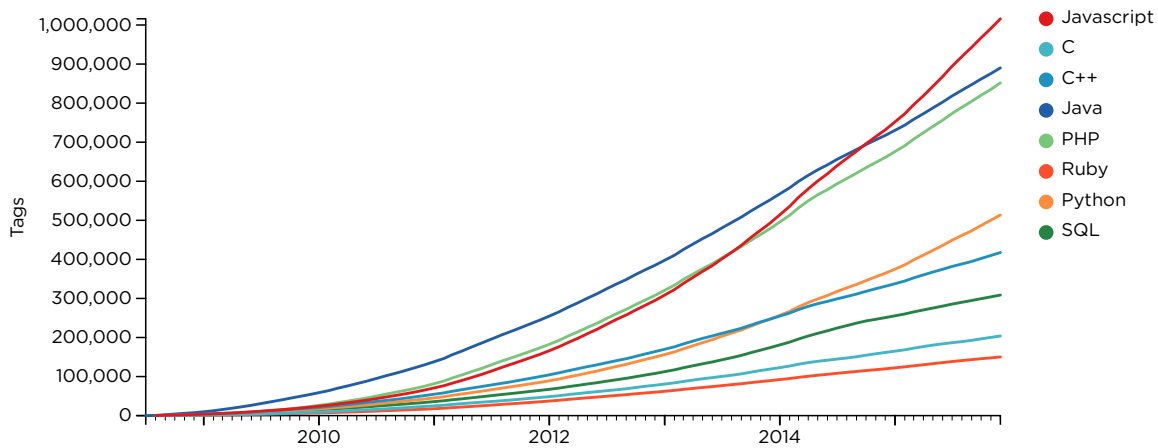


Figure 3.7 - StackOverflow Tags evolution

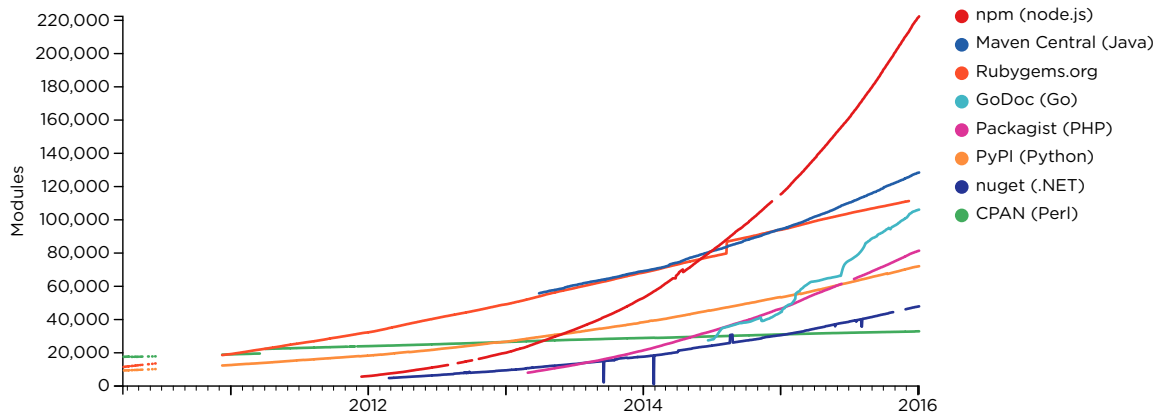


Figure 3.8 - Module Counts per package manager

some additional hints on the situation. Javascript is the third most wanted skill, according to *Indeed*⁷, right after SQL and Java.⁸ Moreover, according to *breaz.io*⁹, Javascript developers get more opportunities than any other developers. Javascript is increasingly adopted in the software industry.

⁷<http://www.indeed.com>

⁸<http://www.indeed.com/jobtrends?q=Javascript%2C+SQL%2C+Java%2C+C%2B%2B%2C+C%2FC%2B%2B%2C+C%23%2C+Python%2C+PHP%2C+Ruby&l=>

⁹<https://breaz.io/>

×

According to these various sources, multi paradigm languages like Javascript are the most widely used by the community and the industry, as presented in table ?? . OOP and Imperative Programming remains strong in the industry, but are progressively abandoned by the community. Functional Programming is not well represented in community nor the industry.

Model	Community support	Industrial need	→	Adoption
Imperative Programming	2	3		2
Object-Oriented Programming	4	5		4
Functional Programming	1	2		1
Multi Paradigm	5	5		5

Table 3.2 – Adoption of Modular Programming Platforms

3.2.3 Efficiency Limitations

Eventually, the presented languages are hitting a wall on their way to performance. They provide global memory abstraction on which to rely to assure encapsulation and composition. Functional programming relies on immutable message-passing. It might impacts performance at a fine-grain level because of heavy memory usage. And the synchronization required by mutable state is often hard to develop with [2], or avoid parallelism [129, 104]. These results are recapped in table 3.3.

Model	Fine-grain level synchronization	Coarse-grain level message passing	→	Efficiency
Imperative Programming	4	1		1
Object-Oriented Programming	4	1		1
Functional Programming	1	4		1
Multi Paradigm	4	1		1

Table 3.3 – Efficiency of Modular Programming Platforms

The only solution to provide performance efficiency is to combine mutable state at a fine-grain level and immutable state at a coarse-grain level ¹⁰. That is to provide both synchronization and message-passing to allow parallelism. The platforms extending these languages with concurrent or parallel paradigms to improve efficiency are addressed in the next section.

3.2.4 Summary

The evaluation of the platforms presented in this section is summarized in table 3.4.

¹⁰<http://joeduffyblog.com/2010/07/11/thoughts-on-immutability-and-concurrency/>

Model	Productivity	Efficiency	Adoption
Imperative Programming	3	1	2
Object-Oriented Programming	5	1	4
Functional Programming	5	1	1
Multi Paradigm	5	1	5

Table 3.4 – Summary of Modular Programming Platforms

3.3 Efficiency Focused Platforms

Both the academia and the industry proposed solutions with efficiency in mind to cope with the limitations the previous section concludes on. The concurrent and parallel programming paradigms are presented in section 3.3.1. The improvement of efficiency impacted the adoption of these platforms, as presented in section 3.3.2. Finally, section 3.3.3 presents the limitations of parallelism on productivity.

3.3.1 Concurrency

Web servers need to be able to process lots of concurrent operations in a scalable fashion. Concurrency is the ability to make progress on several operations roughly simultaneously. It implies to draw memory boundaries to define independent regions, or to define causality in the execution of tasks. When both boundaries and causality are clearly defined, the tasks are independent and can be scheduled in parallel to make progress strictly simultaneously.

The definition of independent tasks allows the fine level synchronization within a task, and coarse level message passing between the tasks required for performance

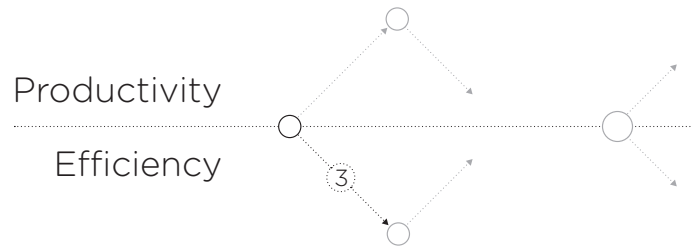


Figure 3.9 – Focus on Efficiency

efficiency. The synchronization of execution at a fine level assures the invariance on the shared state, and avoid communication overhead. The message-passing at a coarser level assures the parallelism. The two are indispensable for efficiency.

3.3.1.1 Concurrent Programming

“
No matter how great the talent or efforts, some things just take time. You can't produce a baby in one month by getting nine women pregnant.”

— Warren Buffett¹¹

Concurrent programming provides the mechanisms to assure atomicity of concurrent operations. They define the causal scheduling of execution and assure the invariance of the global memory. There are two scheduling strategies to execute concurrent tasks on a single processing unit, cooperative scheduling and preemptive scheduling.

Cooperative Scheduling allows a concurrent execution to run until it yields back to the scheduler. Each concurrent execution has an atomic, exclusive access on the memory.

¹¹<http://www.goodreads.com/quotes/476827-no-matter-how-great-the-talent-or-efforts-some-things>

Preemptive Scheduling allows a limited time of execution for each concurrent execution, before preempting it. It assures fairness between the tasks, such as in a multi-tasking operating system. But the unexpected preemption breaks atomicity, the developer needs to lock the shared state to assure atomicity and exclusivity.

The event-driven programming model relies on cooperative scheduling, and the multi-threading programming model relies on preemptive scheduling.

Event-Driven Programming

In the event-driven execution model, developers explicitly define concurrent tasks. The execution model queues and schedules these tasks sequentially to assure exclusivity on shared resources, and cooperatively to assure atomicity. Moreover, they communicate with the resources asynchronously to avoid waiting. A task yields execution to a future task to complete the communication.

Promise were introduced as an help to nicely chain these concurrent tasks [111]. A promise is a placeholder for a future result allowing to defer operations for when the result is available. It layouts the concurrent tasks in chains, similarly to a pipeline.

This execution model is very efficient for highly concurrent applications as it avoids contention. Several platform rely on this execution model, like TAME [104], Node.js¹² and Vert.X¹³. As well as some web servers like Flash [129], Ninja [64] thttpd¹⁴ and Nginx¹⁵.

Though, the event-driven model is limited in performance because of the global memory. The tasks cannot be scheduled in parallel. Lock-free data-structure intends to improve performance by reducing the atomic portions of operations to a minimum.

Lock-Free Data-Structures

The wait-free and lock-free data-structures use atomic operations small enough so that locking is unnecessary [106, 83, 81, 82, 9]. They provide concurrent implementations of basic data-structures such as linked list [165, 159], queue [152, 171], tree

¹²<https://nodejs.org/en/>

¹³<http://vertx.io/>

¹⁴<http://acme.com/software/thttpd/>

¹⁵<https://www.nginx.com/>

[135] and stack [80]. They rely on instructions provided by transactional memories [76] that combine read and write instructions.

Lock-free Data-structures allow parallelization of the concurrent tasks, but are mostly unavailable on common hardware. Multi-threading intends to provide parallelization with software protection.

Multi-Threading Programming

Threads are small execution containers sharing the same memory execution context within an isolated tasks [46], and scheduled in parallel with fork/join instructions [136, 56, 108]. They execute statements synchronously and are preempted to avoid blocking the progression of other threads. Without protection, the preemption breaks the atomicity and exclusivity of memory accesses. To assure atomicity and exclusivity, hence invariance, multi-threading programming models provide protection mechanisms, such as semaphores [43], guarded commands [44], guarded region [75] and monitors [89].

Developers tend to use the global memory extensively, and threads require to protect each and every shared memory cell. This heavy need for synchronization leads to bad performances, and is difficult to develop with [2].

Hybrid Models

Hybrid models join the productivity advantage of synchronous execution from thread, with the efficiency advantage of asynchronous execution from events-driven models. The implementations of hybrid models are libasync [38], InContext [174], Fibers [2], Capriccio [18], Monadic hybrid concurrency [110] and Scala Actors [73]. For example, cooperative threads, or fibers, avoid splitting the execution into atomic tasks nor use protection mechanisms to assure exclusivity. A fiber yields the execution to another fiber to avoid blocking the execution during a long-waiting operation, and recovers it at the same point when the operation finishes. Hiding the yielding points effectively hide the preemption. Developers cannot protect atomicity of operations, which increase the likeliness of race conditions ¹⁶.

¹⁶<https://glyph.twistedmatrix.com/2014/02/unyielding.html>

Limitation of Concurrent Programming

The presented concurrent programming paradigms assure sequentiality of execution within a task and the causal ordering between tasks. Multi-threading imposes heavy protection mechanisms and fails to avoid contention because of the preemptive scheduling. Hybrid models tries to improve over multi-threading using cooperative scheduling. But the synchronous execution imposed by these two models is excessive. It impacts performance, and is difficult to manage efficiently.

The causal ordering between tasks proposed by the event-driven execution model is sufficient to assure correctness of execution [105, 138]. And it is easier for developer to define causal scheduling than to assure the consistency of memory with protection mechanisms. But because of the lack of memory isolation, the concurrent tasks are not scheduled in parallel. It impacts efficiency, as detailed in table 3.5. Yet, parallel programming is the only solution for efficiency, at the expense of development efforts to explicitly define the memory isolation of concurrent tasks and their communications by message passing.

3.3.1.2 Parallel Programming

The Flynn's taxonomy [52] categorizes parallel executions in function of the multiplicity of their flow of instruction and data. Parallel programming models belong to the category Multiple Instruction Multiple Data (MIMD), which is further divided into Single Program Multiple Data (SPMD) [12, 39, 40] and Multiple Program Multiple Data (MPMD) [27, 25]. SPMD defines a single program replicated on many processing units [37, 97, 26] – it is roughly derived from the multi-threading programming model presented above. While MPMD defines multiple parallel tasks in the implementation [65, 54, 53]. The two major theoretical models for parallel programming are the actor model and coroutines.

Actor Model

The Actor model allows to express the causal ordering of computation as a set of parallel actors communicating by asynchronous messages [84, 85, 33]. In reaction to a received message, an actor can create other actors, send messages, and choose how to respond to the next message. Additionally, the communication in reality are too slow compared to execution to be synchronous, and are subject to various faults

and attacks [107]. The Actor model takes these physical limitations in account [86].

Communicating Sequential Processes

Similar works include the Communicating Sequential Processes (CSP) [88, 20], and the Kahn Networks [100, 101]. Similarly, coroutines are autonomous programs which communicate with adjacent modules as if they were input and output subroutines [36]. It defines a pipeline to implement multi-pass algorithms.

3.3.1.3 Summary of Concurrent and Parallel Programming Models

Compared to parallel programming, the concurrent programming models are unable to correctly isolate the concurrent tasks and communicate by message-passing to allow parallelism. It results that the parallel programming model are more efficient than the concurrent programming models, as detailed in table 3.5.

Model	Fine-grain level synchronization	Coarse-grain level message passing	→	Efficiency
Event-driven programming	5	2		2
Lock-free Data-Structures	5	2		2
Multi-threading programming	5	1		1
Hybrid Models	5	1		1
Actor Model	5	5		5
Communicating Sequential Processes	5	5		5

Table 3.5 – Efficiency of Concurrent and Parallel Programming Platforms

3.3.2 Adoption

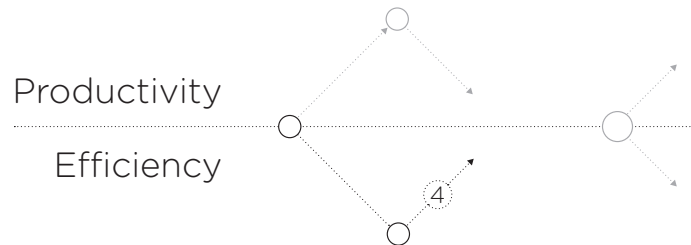


Figure 3.10 – Steering back toward Productivity

When the need for efficiency is higher than the need for productivity, the adoption is steered by the industry more than the community. If the industry really needs a platform, it will commit the required development effort despite a low productivity. The platforms for the Mars Rover or some banking systems are about 30 years old, yet the industry continues to maintain them. The platform presented in this section emerged from the academia and the industry but are often barely known by the larger community of developers. The more the platform trades productivity for efficiency, the less support it will received from the community.

3.3.2.1 Concurrent Programming

Most programming languages implementation supports concurrent programming somehow. Either with multi-threading or event-driven programming. These two are highly adopted by both the industry and the community, as presented in table 3.6.

On the other hand, lock-free data structures and cooperative threads comes from the academia, similarly to fonctionnal programming, and did not encounter significant adoption from the community.

3.3.2.2 Parallel Programming

There exists several platforms directly inspired by the actors model, like Erlang [11, 122, 10], Scala [127], Akka¹⁷ and Play¹⁸. Scala is a programming language unifying the object model and functional programming. Akka is a framework based on Scala, following the actor model to build highly scalable and resilient applications. Play is a web framework based on top of Akka. And Erlang is a functional language designed by Ericsson to operate networks of telecommunication devices [11, 122, 10]

There are as well other platforms inspired by other theoretical model, like Go¹⁹, inspired by Coroutines and CSP. Go is an open source language initiated by Google to build highly concurrent services.

These examples of implementation are largely used in the industry, but are almost unknown outside of it. They are backed by strongly passionate but small communities.

Indeed, it is difficult for developers to manage the superposition of these two organizations, tasks and modules. The organization in independent tasks is hardly compatible with the modular organization presented in the previous section. This superposition makes these platforms accessible only to an elite in the industry supporting it. To mitigate this difficulty, various platforms help organizing the tasks, or adopt the same granularity for modules and tasks to fit the two organizations.

Tasks Organization and Communications

To reduce the difficulties of the superposition of tasks and modules, algorithmic skeletons propose predefined patterns of organization to fit certain types of problems [34, 41, 117, 62]. Developers focus on their problem and delegate the communications to specialized skeletons. These solutions are hardly used by the community, but are crucial in some industrial contexts. A famous example is the map/reduce pattern introduced by Google [41].

Tasks Granularity

The Service Oriented Architectures (SOA) allows developers to express an application as an assembly of services connected to each others. Some examples of

¹⁷<http://akka.io/>

¹⁸<https://www.playframework.com/>

¹⁹<https://golang.org/>

SOA platforms are OSGi²⁰, EJB²¹ and Spring²². It allows to adjust the granularity of tasks to help developers to better fit the tasks organization with the modular organization [1].

More recently, Microservices are tackling the same challenge on the web, with a finer granularity [50, 55, 121]. Some examples of Microservices are Seneca²³. They are very recent, and it is difficult to assess their usage in the community nor the industry. But they seem to be increasingly adopted, both in the industry and in the community.

×

The parallel programming platforms previously presented allow to build generic distributed systems. In the context of the web, a real-time application must process high volumes streams of requests within a certain time. This requirement imposes specific challenges for the platforms.

3.3.2.3 Stream Processing Systems

Data-stream Management Systems

Database Management Systems (DBMS) historically processed large volume of data, and they naturally evolved into Data-stream Management System (DSMS) to processed data streams as well. Because of this legacy, they are in rupture with MPMD platforms presented until now. They borrow the syntax from SQL to run requests in parallel on continuous data streams. The computation of these requests spread over a distributed architecture. Some recent examples are DryadLINQ [95, 176], Apache Hive [158], Timestream [132], Shark [173].

²⁰<https://www.osgi.org/developer/specifications/>

²¹<http://www.oracle.com/technetwork/java/javaee/ejb/index.html>

²²<http://projects.spring.io/spring-framework/>

²³<http://senecajs.org/>

Pipeline Stream Processing

The pipeline architecture organizes an application as a network of event-driven stages connected by explicit queues, the output of one feeding the input of the next [170]. The event-driven paradigm of a stage is similar to work like Ninja [64] and Flash [129] previously presented. But the independence of stages allow to spread the execution on a parallel architecture. The academic works and industrial implementations of pipeline architecture are SEDA [170], StreaMIT [157], Spidle [35], Aspen [164], Pig Latin [128], MEDA [74], CBP [112] and S4 [124] designed at Yahoo, Piccolo [131], Comet [79], Nectar [67], SEEP [118], Legion [15], Spark Streaming [178], Naiad [120] designed at Microsoft, Millwheel [4] designed at Google, Halide [134], Storm [160] by Twitter, SDG [49], Regent [146] and Neptune [21].

×

Parallel programming emerges mainly from industrial needs and academic research but is barely supported by the community. The implementations improve efficiency, but prevent their adoption by the community due to a weak productivity. As summarized in table 3.6, the event-driven programming model is the best candidate for a concurrent programming model regarding adoption. It is supported by the community, and respond to concrete needs in the industry.

Model	Community support	Industrial need	→	Adoption
Event-driven programming	5	5		5
Lock-free Data-Structures	0	2		0
Multi-threading programming	3	5		3
Hybrid Models	0	0		0
Actor Model	1	5		1
Communicating Sequential Processes	1	5		1
Skeleton	2	5		2
Service Oriented Architecture	3	4		3
Microservices	3	3		3

Table 3.6 – Adoption of Concurrent and Parallel Programming Platforms

3.3.3 Productivity Limitations

Parallel programming requires the organization of execution and memory into independent tasks. This independence imposes different granularity of state accessibility. At a fine level, the state is shared, while at a coarser level, it is isolated. It avoids higher-order programming. Hence, it limits the composition of modules and negatively impacts productivity.

Without this composition between modules, parallel programming forces to develop two mental representations – one for the module organization and one for the tasks organization – or to abandon the module organization and productivity altogether. It makes parallel programming accessible only to an elite of developers that are able to be productive despite the two mental representations, as presented in

table 3.7.

Model	Composition	Encapsulation	→	Productivity
Event-driven programming	5	5		5
Lock-free Data-Structures	5	5		5
Multi-threading programming	4	4		4
Hybrid Models	4	4		4
Actor Model	2	2		2
Communicating Sequential Processes	2	2		2
Data Stream System Management	2	3		2
Pipeline Stream Processing	2	3		2

Table 3.7 – Productivity of Concurrent, Parallel and Stream Programming Platforms

3.3.4 Summary

The parallel programming platforms presented in this section focus more on efficiency than productivity. Whereas the concurrent programming platforms are more features better productivity. But no platforms features both productivity and efficiency, as presented in table 3.8. Moreover, except for Event-driven and Multi-threading programming, this lack of versatility impacts adoption.

Model	Productivity	Efficiency	Adoption
Event-driven programming	5	2	5
Lock-free Data-Structures	5	2	0
Multi-threading programming	4	1	3
Hybrid Models	4	1	0
Actor Model	2	5	1
Communicating Sequential Processes	2	5	1
Data Stream System Management	2	5	1
Pipeline Stream Processing	2	5	1

Table 3.8 – Summary of Concurrent and Parallel Programming Platforms

3.4 Adoption Focused Platforms

The platform previously presented focus on productivity or efficiency. The previous section concludes that favoring one negatively impacts the other. Moreover, a balance between productivity and efficiency is required to be both supported by the community and needed by the industry, hence trigger a virtuous circle of adoption. Some platforms feature an abstraction of the tasks organization to allow developers to focus on the modular organization to keep both productivity and efficiency. This abstraction happens either at compile time or at runtime.

3.4.1 Abstraction of Tasks Organization

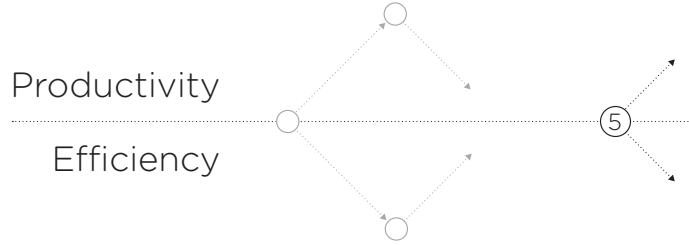


Figure 3.11 – Focus on Adoption

3.4.1.1 Compilers

“*It is a mistake to attempt high concurrency without help from the compiler*”

— R. Behren, J. Condit, E. Brewer [17]

The idea to bridge the gap between the tasks and modules organizations dates from the initial paper that presented this gap [130]. However, the implementation of this idea remains a work in progress.

Parallelism Extraction

To extract parallelism from a sequential implementation, a compiler needs to identify the commutative operations to parallelize their executions [140, 32]. The parallelization of loop iterations has been thoroughly studied [116, 6, 30, 13, 133], particularly with the polyhedral compilation method [14]. Examples of polyhedral compilers are AlphaZ [177], Polly [66] and GRAPHITE [161]. To improve performance gains outside of loops, some compilers identify parallelism in the data-flow representation on the whole program [16, 23, 109].

Data processing applications [49] such as web services [142] are often already roughly organized as data-flow. In higher-order programming, continuous passing

style and promises encourage this data-flow organization. However, the mutable closures required for higher-order programming remains a challenge to parallelize because they rely on a globally shared memory [77, 125, 115]. To extract parallelism, compilers rely on static analysis or notations from the developers.

Static analysis

Compilers analyze the source code of a program to detect commutative operations in the control flow [5]. The point-to analysis is a method for static analysis. It identifies aliasing, that is multiple symbolic names pointing to the same memory location, hence identifies side-effects [8, 96, 148, 169] which allows to infer commutativity. Another method, the abstract interpretation, is to interpret the possible path of executions with virtual inputs. It allows to statically reason on the behavior of dynamic programs [113, 147, 58, 72, 137, 57, 19]. It is successfully used for security applications to detect malicious scripts, or obfuscate code [91, 99, 175, 114, 31, 47]

However, these static analysis techniques remains often too imprecise, and expensive for the performance gain to be profitable in dynamic languages, such as Javascript [144]. Instead, some compilers relies on annotations from the developers.

Annotations

Some works proposed to rely on annotations from the developer to identify the shared data structures and infer the commutativity of operations [166, 49]. Such annotations are especially relevant for accelerators such as GPUs or FPGAs, because the development effort yields huge performance improvements [155]. Examples of such compilers are .

Compilation Limitations

For dynamic languages like Javascript, the static analysis is not sufficient to correctly infer the independence of tasks to parallelize them. Parallel compilers often fall back on relying on annotation provided by developers. Hence, the burden of detailing the tasks and memory organizations falls back to the developer. It impacts productivity and adoption.

3.4.1.2 Runtimes

At runtime, the incertitudes on the independence of tasks is resolved. It allows analyzes precise enough to detect and distribute the commutative operations.

Partitioned Global Address Space

The Partitioned Global Address Space (PGAS) provides a uniform access on a distributed memory architecture. It attempts to combine the efficiency of distributed memory systems, with the productivity of shared memory systems. Each computing node executes the same program, and provide its local memory to be shared with all the other nodes. The PGAS platform assures the remote accesses and synchronization of memory across nodes. Examples of implementation of the PGAS model are CoArray Fortran [126], X10 [29], Unified Parallel C [60], Chapel[24], OpenSHMEM [28], Kokko [48], UPC++ [179], RAJA [90], ACPdl [3], HPX [102, 103].

Dynamic Distribution of Execution

Following SEDA, Leda proposes a model where the independent stages of the pipeline are defined only by their role in the application [142, 143]. The execution distribution and module organization are different. The actual execution distribution is defined automatically during deployment. This automation manages the execution organizations to help the developer focus on the modular organization.

3.4.1.3 Productivity and Efficiency

The platforms presented in this section intend to merge both productivity and efficiency in a single platform by bringing parallelism to productivity languages. Because of this base, they feature decent encapsulation, but they limit the use of higher-order programming between tasks to allow their isolation. Hence, it degrades composition, as presented in table 3.9.

Despite worse productivity, this parallelization bring good efficiency, as presented in table 3.11.

Model	Composition	Encapsulation	→	Productivity
Polyhedral Compilers	2	4		2
Annotation Compilers	3	4		3
Partitioned Global Address Space	2	4		2
Dynamic Distribution	2	4		2

Table 3.9 – Productivity of Compilation and Runtime Platforms

Model	Fine-grain level synchronization	Coarse-grain level message passing	→	Efficiency
Polyhedral Compilers	5	4		4
Annotation Compilers	5	4		4
Partitioned Global Address Space	4	4		4
Dynamic Distribution	4	4		4

Table 3.10 – Efficiency of Compilation and Runtime Platforms

3.4.2 Limitation

The platforms presented in this section come from the need of the industry to reduce the development commitment required for efficiency. However, these platforms re-

spond exclusively to academic or industrial needs, and are barely supported by the community, as presented in table 3.11.

The balance between efficiency and productivity is not sufficient for a community of passionate to arise. The platforms need to allow novice to start learning and experimenting at small scale. It incites the community to start projects, and grow them organically to build businesses. The context of web development is particularly adapted for this growth.

Model	Community support	Industrial need	→	Adoption
Polyhedral Compilers	0	3		0
Annotation Compilers	2	3		2
Partitioned Global Address Space	1	3		1
Dynamic Distribution	0	1		0

Table 3.11 – Adoption of Compilation and Runtime Platforms

3.4.3 Summary

As detaild in table 3.12, this section proved that a platform cannot trade productivity against efficiency without massively losing the community required to trigger the adoption.

Model	Productivity	Efficiency	Adoption
Polyhedral Compilers	2	4	0
Annotation Compilers	3	4	2
Partitioned Global Address Space	2	4	1
Dynamic Distribution	2	4	0

Table 3.12 – Summary of Compilation and Runtime Platforms

3.5 Discontinuous Developments

This chapter presented a broad view of platforms and their balance between productivity and efficiency. It established that the platforms favoring one eventually sacrifice the other. Moreover, the adoption of these platforms proves that none of these compromises are sustainable. Indeed, as presented in table 3.13, no platforms provides productivity, efficiency and adoption.

Model	Productivity	Efficiency	Adoption
Imperative Programming	3	1	2
Object-Oriented Programming	5	1	4
Functional Programming	5	1	1

Multi Paradigm	5	1	5
Event-driven programming	5	2	5
Lock-free Data-Structures	5	2	0
Multi-threading programming	4	1	3
Hybrid Models	4	1	0
Actor Model	2	5	1
Communicating Sequential Processes	2	5	1
Data Stream System Management	2	5	1
Pipeline Stream Processing	2	5	1
Polyhedral Compilers	2	4	0
Annotation Compilers	3	4	2
Partitioned Global Address Space	2	4	1
Dynamic Distribution	2	4	0

Table 3.13 – Summary of the state of the art

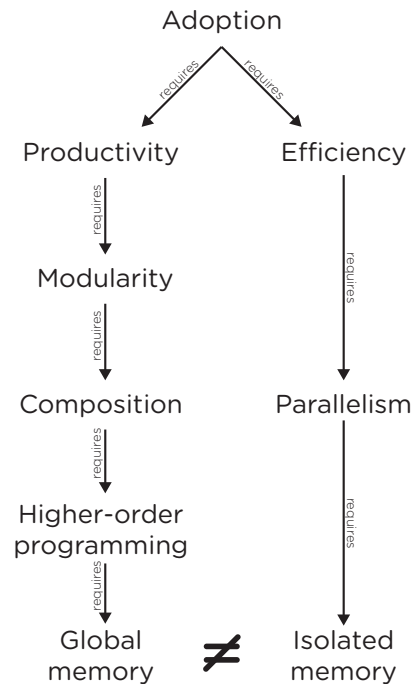
×

This chapter highlighted that productivity requires modularity through encapsulation and composition. Productivity requires higher-order programming which relies on a global memory abstraction as explained in section 3.2.3. Whereas efficiency requires a balance between fine-grain level shared state with synchronization and coarse-grain level independence with message-passing. This discontinuity between fine-grain level and coarse-grain level avoids the global memory abstraction, hence productivity. The absence of a global memory abstraction reserves efficient

platforms for an elite of developers. No platform can support simultaneously productivity and efficiency. Nonetheless, a platform needs to be adopted both by the industry and the community to be sustainable.

The conclusion of this chapter is that no platform is able to follow a project from the early beginning until the industrial maturation of the project. All the platforms tend to be stucked in a compromise between these two goals, and cannot follow the evolution required for this compromise. These compromises are rigidly defined, while the need of the application is constantly evolving. They lack the possibility to follow the organic evolution of a project. Therefore, a project needs to change platform to change its priority, which leads to economical consequences.

To avoid these consequences, platforms would need to support productivity to allow the community to experiment, and organically start projects. And then continuously shift toward efficiency as the project evolves, and requires it.



CHAPTER 4

SEAMLESS SHIFT FROM PRODUCTIV- ITY TO EFFICIENCY

Contents

4.1 Proposition	68
4.1.1 Continuous Development	69
4.1.2 Equivalence	69
4.1.2.1 Execution Independence	69
4.1.2.2 Rupture points	70
4.1.2.3 Invariance	71
4.1.2.4 Transformation	72
4.2 Execution Models	74
4.2.1 Event-Driven Execution Model	74
4.2.1.1 Continuation Passing Style	75
4.2.1.2 Promise	76
4.2.2 Fluxional Execution Model	77
4.2.2.1 Fluxion	78
4.2.2.2 Streams	79
4.2.3 Examples	79
4.2.3.1 Fluxional Application	79
4.2.3.2 Transformation	81
4.3 Evaluation	83
4.3.1 Productivity and Adoption	84
4.3.2 Efficiency	84
4.3.3 Summary	85

The evolution of the economical constraints of a web application requires to continuously shift from productivity to efficiency. The incompatibility between the two organizations forces platforms to propose a compromise between the two. It makes it impossible for these platforms to follow the evolution of a web application. Hence, it implies technological ruptures during this evolution. Huge developing efforts are pulled to translate manually from one organization into the other, and later to maintain the implementation.

4.1 Proposition

This thesis proposes a platform allowing a seamless shift of focus to follow the development of a web application from the productivity required in the early beginning until the efficiency required during maturation.

Developers start a project without compromises on productivity. The language targets an event-driven execution model. And they continuously transform their implementation to target the more efficient pipeline architecture.

Node.js implements Javascript, a productivity language, with an event driven execution model to implement web applications with decent performances. However, the performance of this execution model is limited by the sequentiality of execution required to preserve exclusivity of memory accesses. The pipeline execution model overstep this performance limitation. It enforces memory isolation between stages allowing the parallel execution required for efficiency. But this isolation limits the productivity because it avoids higher-order programming. The difference in the memory abstractions between the two execution models is illustrated in figure 4.1.

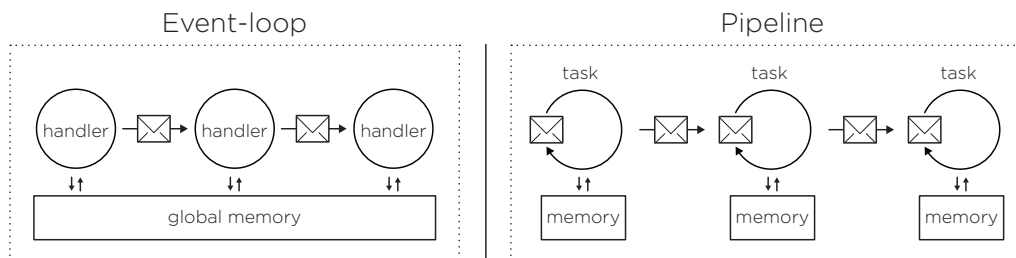


Figure 4.1 – Differences of memory abstraction

Despite this difference, these two execution models present an interesting simi-

larity. They both organize the execution as a sequence of tasks causally scheduled. This thesis proposes an equivalence between these two execution models based on this similarity. Following this equivalence, it proposes a compiler that distributes the global memory into isolated stages of the pipeline. Concretely, it transforms a mono-threaded, event-driven application to run on a parallel pipeline architecture.

4.1.1 Continuous Development ---

This transformation allows a continuity of compromises between productivity and efficiency to seamlessly follow the shift of focus during development.

At first, the focus remains on the productivity of development rather than the efficiency of execution. The development begins with the event-driven model to take advantage of the productivity of the global memory abstraction. The execution resulting from the transformation is as efficient as the original event-driven execution model.

During the maturation of the application, the focus continuously shift towards efficiency. The transformation distribute the global memory into isolated stages as much as possible. It allows developer to identify the dependencies in this global memory avoiding the distribution. They can identify these dependencies, and arrange the implementation accordingly to allow parallelism. It helps developers to enforce efficiency through continuous iteration, instead of disruptive shifts of technology.

4.1.2 Equivalence ---

The equivalence between these two execution models is broken down in two steps. The first step identifies the stages in the control flow, by detecting rupture points between them. The second step enforces isolation of memory between these stages, and replaces synchronization with message passing to preserve the invariance.

4.1.2.1 Execution Independence

In the pipeline architecture, each stage has its own thread of execution independent from the others. Whereas in the event-driven execution model, the handlers are executed sequentially. Despite this difference, the execution of a handler is as independent as the execution of a stage of a pipeline. The call stacks of two handlers

are isolated, as illustrated in figure 4.2. Indeed, a handler holds the execution until its call stack is empty, when all synchronous function calls terminates. Only then it yields the execution to the event-loop, which schedule the next handler. A rupture point separates the call stacks of two handlers.

For simplification, the figures 4.2 illustrates a rupture point with only two interleaving chains of handlers, but there could be many more chains interleaving. Each chain is comparable to a thread in multi-thread programming. The two chains in figure 4.2 are $\rightarrow \textcircled{1} \rightarrow \textcircled{2} \rightarrow$ and $\rightarrow \textcircled{3} \rightarrow \textcircled{4} \rightarrow$. And in figures 4.5, 4.6, ?? and ??, they are $\rightarrow \textcircled{1} \rightarrow \textcircled{3} \rightarrow$ and $\rightarrow \textcircled{2} \rightarrow \textcircled{4} \rightarrow$.

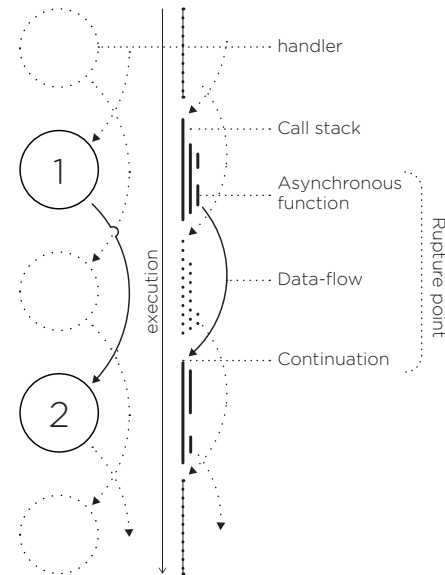


Figure 4.2 - Rupture point

4.1.2.2 Rupture points

A rupture point is a call of a loosely coupled function. It is an asynchronous call without subsequent synchronization with the caller. This asynchronous function call is equivalent to a data-flow between two stages in the pipeline architecture. The parent sends a message to the child handler containing the result of the asynchronous function call it initiated. The event-driven execution model expects callback to send the message and continue the execution once the asynchronous call completes.

Callbacks, Listeners and Continuations

A callback is a function passed as a parameter to a function call. It is not inherently asynchronous. Only two kinds of callbacks are asynchronous, listeners and continuations. They are invoked to continue the execution with data not yet available synchronously, in the caller context. Listeners listen to a stream of events, hence are invoked multiple times. Whereas continuations are invoked only once to continue the execution after an asynchronous operation completes. The two corresponding types of rupture points are *start* and *post*.

Start rupture points (listeners) are on the border between the application and the outside, continuously receiving incoming user requests. An example of a start

rupture point is in listing 4.5, between the call to `app.get()`, and its listener handler. These rupture points indicate the input of a data stream in the program, and the beginning of a chain of fluxions to process this stream.

Post rupture points (continuations) represent a continuity in the execution flow after an asynchronous operation yielding a unique result, such as reading a file, or a database. An example of a post rupture points is in listing 4.5, between the call to `fs.readFile()`, and its continuation `reply`.

The detection of rupture points allows to retrieve the data-flow and the stages for a pipeline architecture from the implementation following the event-loop model. The implementation of this detection is fully addressed in the next chapter, in sections 5.1.3 and 5.2.1. However, these stages still require a global memory to communicate. They can't be executed in parallel without breaking the invariance.

4.1.2.3 Invariance

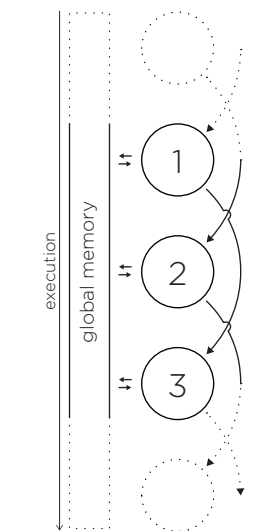


Figure 4.3 - Sequential scheduling

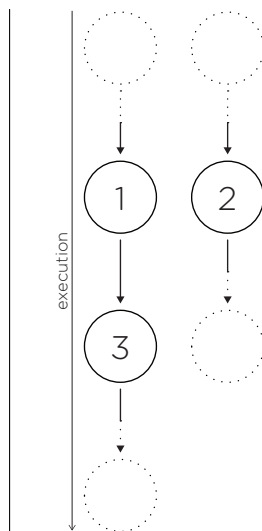


Figure 4.4 - Causal scheduling

A global memory requires the sequential scheduling of handlers to assure exclusivity of access, as illustrated in figure 4.5. The global memory is the only reason for this sequential execution. Yet, the causal scheduling of tasks is sufficient to assure the correctness of the execution.

Message passing only requires causal scheduling of handlers which allows parallelism, as illustrated in figure 4.6. If the handlers didn't rely on the global memory, they could be executed in parallel, as long as their causalities are respected. Following some rules, it is possible to replace their global memory usage with message-passing, and parallel their execution.

4.1.2.4 Transformation

Rule #1 Scope

If an handler uses a variable between the reception of two data, then it needs to store it independently of the global memory. The reliance on this memory imposes the handler to not be reentrant. There cannot be several instances of its execution. The stream of incoming data must be processed sequentially.

Rule #2 Stream

If two handlers causally related rely on the same memory region, the causal relation assures the exclusivity of access. Therefore, the global memory can be replaced by sending the updated memory in the data-flow. As illustrated in figure ??, the upstream handler ① communicates the memory update to the downstream handler ③. And each handler has access only to its own memory.

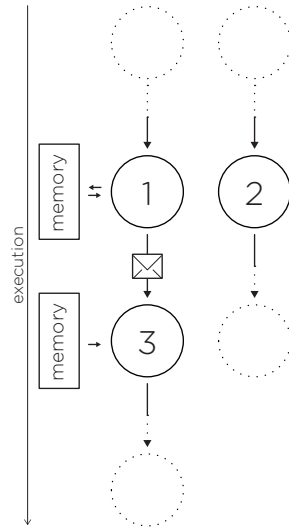


Figure 4.5 - Message passing memory update

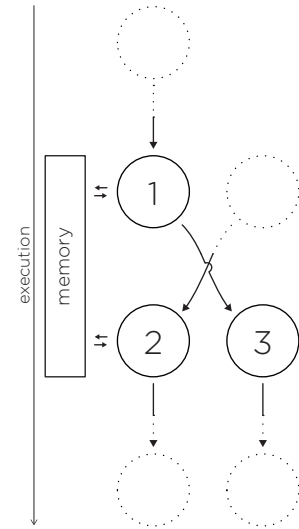


Figure 4.6 - Sequential execution

Rule #3 Share

However, if the downstream handler modifies this memory, it is not possible to isolate it. The upstream handler cannot be notified in time of this modification. Indeed, the upstream handler might already be processing the next datum in the stream. Moreover, if two handlers not causally related rely on the same memory region, they can access it in any order. They need to be scheduled sequentially to maintain the exclusivity

of access, as illustrated in figure ??.

×

By distributing the global memory following these rules, the sequential scheduling can be loosened while preserving invariance to parallelize the execution. This distribution - hence the parallelization - only depends on the memory dependencies between handlers. Of course, at first, the dependencies will remain tight as the focus is on productivity. But, developers can continuously iterate on implementation to loosen the dependencies and improve efficiency.

The implementation of the distribution of the global memory is fully addressed in next chapter, in section 5.2.2.

From Event-loop to Pipeline

Concretely, this transformation turns handler from the event-driven execution models into stages of the pipeline, as illustrated in figure 4.7. And it distributes the global memory into these different stages, as illustrated in figure 4.8. The details of these two execution models important for this transformation are presented in the next section.

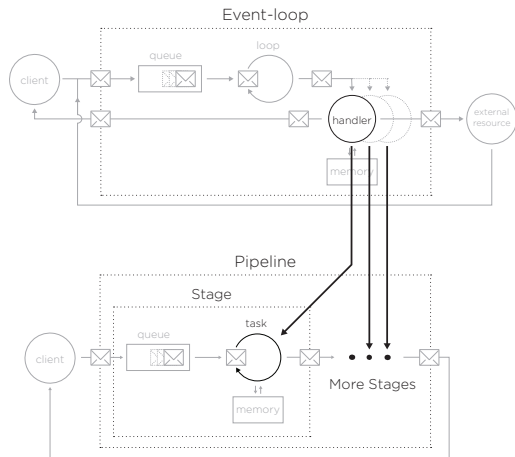


Figure 4.7 - Equivalence between handlers and tasks

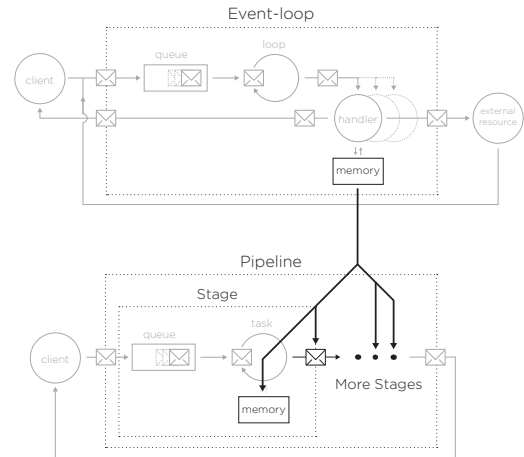


Figure 4.8 - Distribution of the global memory abstraction with message passing

4.2 Execution Models

The event-driven execution model and the pipeline execution model were already briefly presented in chapter 2, section 2.1.2, page 19. The next paragraphs dive into the details of each execution model in regard of the transformation.

4.2.1 Event-Driven Execution Model

The event-driven execution model is processes a queue of asynchronous events by cooperatively scheduling handlers. To respond to an event, the associated handler can directly respond to the source of the event. Or it can request an external resource, and chain another han-

dler to later process the initial event with the resource response, as illustrated in figure 4.9. The developer defines each handler as a continuation and defines their causality using the continuation passing style [168, 78].

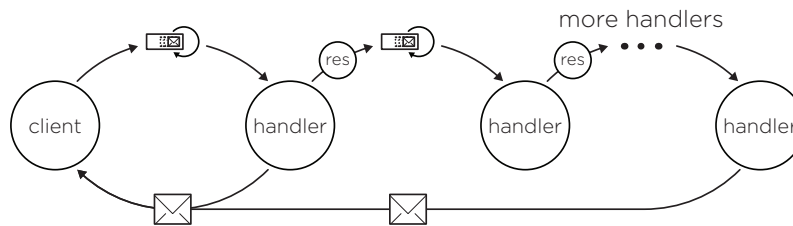


Figure 4.9 - Chain of continuations

4.2.1.1 Continuation Passing Style

A continuation is a function passed as an argument to a function call. The continuation is invoked after the completion of the callee, to continue the execution. In the event-driven execution model, the continuation is invoked as a new handler, to avoid blocking the caller until its completion. At its invocation, the continuation retrieves both the caller context and the result of the callee through a closure. Listing 4.1 illustrates an example of continuation in *Node.js*.

```

1 callee(input
  , function continuation(error, result) {
2   if (error)
3     throw error;
4   // ... modify result
5   console.log(result);
6 });

```

Listing 4.1 - Example of a continuation

Asynchronous continuations cannot be composed to chain their execution like synchronous functions, as illustrated in listing 4.2. This nested construction is sometime difficult to follow. Promises improve continuations to allow this composition. They allow to arrange a sequence of asynchronous operations in a chain, similar to a pipeline.

```

1
2 callee(input
3   , function continuation(error, result) {
4     if (error)
5       throw error;
6     // ... modify result
7     nestedCallee(result, function
8       continuation(error, nestedResult) {
9         if (error)
10           throw error;
11         // ... modify result
12         superNestedCallee
13         (nestedResult, function
14           continuation(error, superNestedResult) {
15             // ... and so on ...
16           }
17         });
18   });
19 });

```

Listing 4.2 – Example of a sequence of continuations

4.2.1.2 Promise

A Promise is used as a placeholder for the eventual outcome of a deferred (and possibly asynchronous) operation. Listing 4.3 illustrate a simple promise. In Javascript, promises expose a **then** method which expects a continuation to continue the execution with the result of the deferred operation. This method allows to chain Promises one after the other, as illustrated in listing 4.4.

```

1 var promise = callee(input)

```



```

2
3 promise.then(function onSuccess(result) {
4   console.log(result);
5 }, function onError(error) {
6   throw error;
7 });

```

Listing 4.3 – Example of a Promise

```

1 callee_promise_1(input)
2 .then(callee_promise_2, onError)
3 .then(callee_promise_3, onError)
4 .then(console.log, onError);
5
6 function onError(error) {
7   throw error;
8 }

```

Listing 4.4 – Example of a chain of Promises

Promises allow to easily arrange the execution flow in parallel or in sequence according to the required causality. Programmers are encouraged to arrange the computation as series of steps to process incoming events and yield outgoing events. In this sense, Promises are an intermediate step toward the pipeline execution model.

4.2.2 Fluxional Execution Model

The fluxional execution model is inspired by the pipeline architecture. It is the target for the transformation presented in this thesis. It executes programs written in the fluxional language, whose grammar is presented in figure 4.10. It intends to provide scalability to web applications with a granularity of parallelism at the function level.

The functions of an application $\langle \text{program} \rangle$ are encapsulated in autonomous execution containers named *fluxions* $\langle \text{flx} \rangle$.

$\langle \text{program} \rangle$	\models	$\langle \text{flx} \rangle \mid \langle \text{flx} \rangle \text{ eol } \langle \text{program} \rangle$
$\langle \text{flx} \rangle$	\models	$\text{flx } \langle \text{id} \rangle \langle \text{tags} \rangle \langle \text{ctx} \rangle \text{ eol } \langle \text{streams} \rangle \text{ eol } \langle \text{fn} \rangle$
$\langle \text{tags} \rangle$	\models	$\& \langle \text{list} \rangle \mid \text{empty string}$
$\langle \text{streams} \rangle$	\models	$\text{null} \mid \langle \text{stream} \rangle \mid \langle \text{stream} \rangle \text{ eol } \langle \text{streams} \rangle$
$\langle \text{stream} \rangle$	\models	$\langle \text{type} \rangle \langle \text{dest} \rangle [\langle \text{msg} \rangle]$
$\langle \text{dest} \rangle$	\models	$\langle \text{list} \rangle$
$\langle \text{ctx} \rangle$	\models	$\{ \langle \text{list} \rangle \}$
$\langle \text{msg} \rangle$	\models	$[\langle \text{list} \rangle]$
$\langle \text{list} \rangle$	\models	$\langle \text{id} \rangle \mid \langle \text{id} \rangle , \langle \text{list} \rangle$
$\langle \text{type} \rangle$	\models	$>> \mid ->$
$\langle \text{id} \rangle$	\models	<i>Identifier</i>
$\langle \text{fn} \rangle$	\models	<i>Source language with $\langle \text{stream} \rangle$ placeholders</i>

Figure 4.10 – Syntax of a high-level language to represent a program in the fluxional form

4.2.2.1 Fluxion

A *fluxion* $\langle \text{flx} \rangle$ is named by a unique identifier $\langle \text{id} \rangle$ to receive messages, and might be part of one or more groups indicated by tags $\langle \text{tags} \rangle$. A *fluxion* is composed of a processing function $\langle \text{fn} \rangle$, and a local memory called a *context* $\langle \text{ctx} \rangle$.

At a message reception, the *fluxion* modifies its *context*, and sends messages to downstream *fluxions* on its output streams $\langle \text{streams} \rangle$. The *context* stores the state on which a *fluxion* relies between two message receptions, similarly to the actors model. The messaging system queues the output messages for the event loop to process them later by calling the downstream *fluxions*.

In addition to message passing, the execution model allows *fluxions* to com-

municate by sharing state between their *contexts*. The fluxions that need this synchronization are grouped with the same tag, and loose their independence.

4.2.2.2 Streams

There are two types of streams, *start* and *post*, which correspond to the nature of the rupture point producing the stream. A *start* rupture point starts a chain of continuations, while a *post* rupture point is a continuation in a chain. *Start* rupture points are indicated with a double arrow ($\rightarrow\!\!\rightarrow$ or \gg) and *post* rupture points with a simple arrow (\rightarrow or \rightarrow).

The variables defined before the *start* are available for the whole chain, and require synchronization for concurrent execution of the same chain. Whereas the variables defined inside the chain are available only in this chain, and don't require synchronization. The two types and implications of rupture points are further detailed in section 5.2.1.

4.2.3 Examples _____

4.2.3.1 Fluxional Application

To test fluxional execution model, it was used to develop a simple web application. This application doesn't use the transformation, nor the continuous development advocated in this thesis. Nonetheless, it is related here to illustrate the basic functioning of the fluxional execution model.



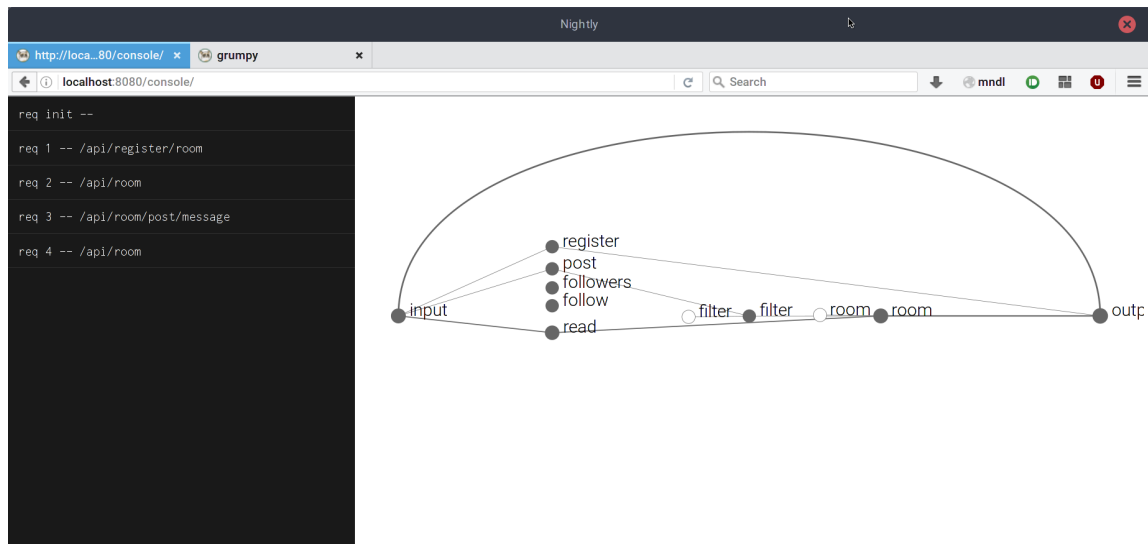


Figure 4.11 – Screenshot from the grumpy console

The application is accessible at <http://grumpy.etnbrd.com>. It allows to post anonymous grumbles into chat rooms. Additionally, a room can subscribe to the grumbles from another room.

The network of fluxions structuring the applications can be monitored at <http://grumpy.etnbrd.com/console>. It is organized into five stages, traversed by the stream of request from left to right. The first and stages are the input and output, connecting the application with the clients. The second stage contains the fluxions receiving and formatting the request, before passing them to the next layer. The fluxion in the third stage is a simple filter before posting grumbles. It is an example of a fluxion modifying a message from the stream. Finally, each room is instantiated as a new fluxion in the fourth stage, storing the received grumbles in its context.

The left panel logs the requests received by the application, and the path of each request can be traced from stage to stage.

The socket descriptor is stuck with the network interface, hence, it cannot be serialized to flow from one fluxion to another. Instead, the first and last fluxions share their memory, and the request flows with an id to associate it with the socket descriptor.

This example application was developed before the implementation of the isolation of fluxions in the fluxional execution model. So it was possible to share their memory by simply sending a memory reference between two fluxions, without grouping them. This explains the direct link between the input and output fluxions.

The example illustrated in the next paragraphs explains in more details the transformation, and the grouping required for fluxion sharing memory. And it explains the impact of the memory dependencies on parallelism and scalability.

4.2.3.2 Transformation

The transformation from continuation passing style to the fluxion execution model is now illustrated with a simple web application.

```

1 var app = require('express')(),
2   fs = require('fs'),
3   count = 0;
4
5 app.get('/', function handler(req, res){
6   fs.readFile(__filename, function reply(err, data) {
7     count += 1;
8     res.send(err || template(count, data));
9   });
10 });
11
12 app.listen(8080);

```

Listing 4.5 - Example web application

The example application in listing 4.5 reads a file, and sends it back along with a request counter. The **handler** function, line 5 to 10, receives the input stream of requests. The **count** variable at line 3 counts the requests, and needs to be saved between two messages receptions. The **template** function formats the output stream to be sent back to the client. The **app.get** and **res.send** functions, lines 5 and 8, interface the application with the clients. Between these two interface functions is a chain of three functions to process the client requests : **app.get** \rightarrow **handler** \rightarrow **reply**. This chain of functions is transformed into a pipeline, expressed in the high-level fluxional language in listing 4.6. The transformation process between the source and the fluxional code is explained in chapter 5, in section 5.2.1.

```

1 flx main & grp_res
2 » handler [res]
3   var app = require('express')(),
4     fs = require('fs'),
5     count = 0;
6
7   app.get('/', » handler); //
8   app.listen(8080);
9
10 flx handler
11 -> reply [res]
12   function handler(req, res) {
13     fs.readFile(__filename, -> reply); //

```

```

14 }
15
16 flx reply & grp_res {count, template}
17 -> null
18 function reply(error, data) {
19   count += 1; //
20   res.send(err || template(count, data)); //
21 }

```

Listing 4.6 – Example application expressed in the high-level fluxional language

The execution is illustrated in figure 4.12. The dashed arrows between fluxions represent the message streams as seen in the fluxional application. The plain arrows represent the operations of the messaging system during the execution. These steps are indicated by numeroted circles. The *program* registers its fluxions in the messaging system, ①. The fluxion *reply* has a context containing the variable **count** and **template**. When the application receives a request, the first fluxion in the stream, *main*, queues a **start** message containing the request, ②. This first message is to be received by the next fluxion *handler*, ③, and triggers its execution, ④. The fluxion *handler* sends back a message, ⑤, to be enqueued, ⑥. The system loops through steps ③ to ⑥ until the queue is empty. This cycle starts again for each new incoming request causing another **start** message.

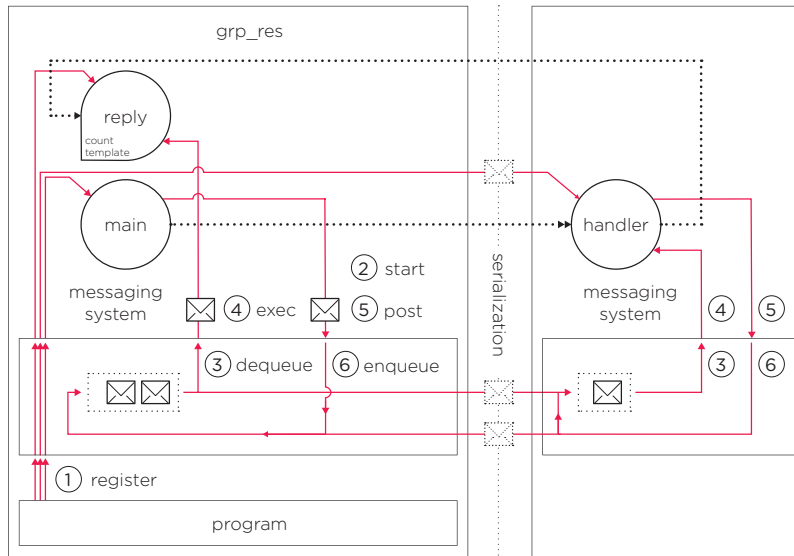


Figure 4.12 – The fluxional execution model in details

The chain of functions from listing 4.5 is expressed in the fluxional language in listing 4.6. The fluxion *handler* doesn't have any dependencies, so it can be

executed in a parallel event-loop. The fluxions `main` and `reply` belong to the group `grp_res`, indicating their dependency over the variable `res`. The group name is chosen arbitrarily. All the fluxions inside a group are executed sequentially on the same event-loop, to protect the shared variables against concurrent accesses.

The variable `res` is created and consumed within a chain of *post* stream. Therefore, it is exclusive to one request and cannot be propagated to another request. It doesn't prevent the whole group from being replicated. However, the fluxion `reply` depends on the variable `count` created upstream the *start* stream, which prevents this replication. If it did not rely on this variable, the group `grp_res` would be stateless, and could be replicated to cope with the incoming traffic.

×

This execution model allows to parallelize the execution of an application as a pipeline, as with the fluxion `handler`. And some parts are replicated, as could be the group `grp_res`. This parallelization improves the efficiency of the application. Indeed, as a fluxion contains its state and expresses its dependencies, it can be migrated. It allows to adapt the number of fluxions per core to adjust the resource usage in function of the desired throughput.

Yet, the parallelization is limited by the dependencies between fluxions. A developer can ignore these dependencies at first, to focus on productivity. And then continuously tune the implementation to remove these dependencies and improve efficiency. This continuous tuning avoids the disruptive shifts of technology required by current platforms.

CHAPTER 5

IMPLEMENTATION

Contents

5.1	Dues	88
5.1.1	Definition	89
5.1.1.1	Usage	89
5.1.1.2	Creation	89
5.1.1.3	Composition	90
5.1.2	From Continuations to Dues	90
5.1.2.1	Execution order	91
5.1.2.2	Execution linearity	91
5.1.2.3	Variable scope	92
5.1.3	Due Compiler	93
5.1.3.1	Identification of continuations	93
5.1.3.2	Generation of chains	94
5.1.3.3	Evaluation	95
5.2	Fluxions	101
5.2.1	Fluxions Compiler	101
5.2.1.1	Detection	102
5.2.2	Fluxions Isolation	103
5.2.2.1	Scope	103
5.2.2.2	Stream	104
5.2.2.3	Share	104
5.2.3	Real test case	105
5.2.3.1	Compilation	106
5.2.3.2	Isolation	107

The transformation allowed by the equivalence from an event-driven program into a distributed network of fluxions is implemented incrementally into two compilers, as presented in figure 5.1. Each compilers is divided into two steps, the identification of the rupture points separating the stages of the pipeline, and the isolation of these stages.

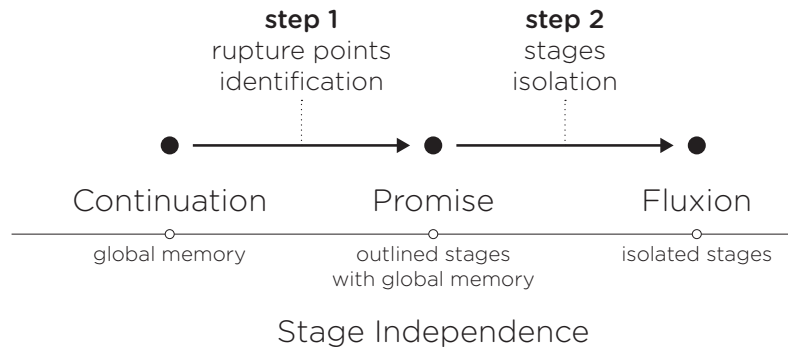


Figure 5.1 - Roadmap

The first compiler focus on the identification of simple chains of causality between continuations to transform these chains into Promises. However, promises are more expressive than the simple chaining of causal sequentiality. Moreover, they impose a different convention than continuations on how to hand back the outcome and errors of the deferred computation. This difference brings unnecessary complexity to the identification of chains. To rule out this difference between continuations and Promises, section 5.1 introduces a simpler specification to Promise, called Due.

The second compiler detects all the chains of causality between continuations and encapsulate them in fluxions. It isolates the fluxions when possible to allow the parallelism required for efficiency. The second compilers is introduced in section 5.2.

5.1 Dues

5.1.1 Definition

A Due is an object used as placeholder for the eventual outcome of a deferred operation. They are essentially similar to ECMAScript Promises¹, except for the convention to hand back outcomes. They use the *error-first* convention, like *Node.js*, as illustrated line 5 in listing 5.1. The implementation of Dues and its tests are available online².

5.1.1.1 Usage

```
1 var my_fn_due = require('due').mock(my_fn);
2
3 var due = my_fn_due(input);
4
5 due.then(function continuation(error, result) {
6   if (!error) {
7     console.log(result);
8   } else {
9     throw error;
10  }
11 });
```

Listing 5.1 – Example of a due

In listing 5.1, the function `my_fn_due` synchronously returns a due as a placeholder for its outcome. The `then` method of the due allows to define a continuation to continue the execution after retrieving the outcome, like line 5. If the deferred operation is synchronous, the Due settles during its creation and the `then` method immediately calls this continuation. If the deferred operation is asynchronous, this continuation is called during the Due settlement.

5.1.1.2 Creation

```
1 Due.mock = function(my_fn) {
2   return function mocked_fn() {
3     var _args = Array.prototype.slice.call(arguments),
4         _this = this;
5
6     return new Due(function(settle) {
7       _args.push(settle);
8       my_fn.apply(_this, _args);
9     })
10  }
11 }
```

¹<http://www.ecma-international.org/ecma-262/6.0/#sec-promise-objects>

²<https://www.npmjs.com/package/due>

Listing 5.2 – Creation of a due

In listing 5.1, line 1, the `mock` method wraps the original function `my_fn` in a Due-compatible function `mocked_fn`. The `mock` method is detailed in listing 5.2 to illustrate the creation of a Due. It returns a Due compatible function, `mocked_fn`, line 2. That is a function that returns a Due, instead of expecting a continuation.

At the execution of `mocked_fn` the Due to be returned is created line 6, with the original function passed as argument. The original function `my_fn` is executed during the creation of the Due. The `settle` function provided is passed as a continuation line 7 for the original function to settle the returned Due. When the original function completes, it calls `settle` to settle the Due and save the outcome. This outcome can then be retrieved with the continuation provided by the `then` method.

5.1.1.3 Composition

Dues arrange the execution flow as a chain of actions to carry on inputs. The composition of Dues in a chain is illustrated in listing 5.3. It is similar to the composition of Promises explained in the previous chapter, section 4.2.1.2, page 76.

```

1 var Due = require('due');
2
3 var my_fn_due_1 = Due.mock(my_fn_1),
4     my_fn_due_2 = Due.mock(my_fn_2),
5     my_fn_due_3 = Due.mock(my_fn_3);
6
7 my_fn_due_1(input)
8   .then(my_fn_due_2)
9   .then(my_fn_due_3)
10  .then(console.log);

```

Listing 5.3 – Dues are chained like Promises

The `then` method of the current Due returns an intermediary Due that settles when the Due returned by the passed continuation settles. As example, in listing 5.3 the Due returned by the `then` method line 8 settles when the Due returned by `my_fn_due_2` settles. It allows to chain continuations one after the other like a pipeline, instead of the nested composition of continuations.

5.1.2 From Continuations to Dues _____

The equivalence between continuations and Dues allows the transformation of a nested imbrication of continuations into a chain of Dues. To preserve the seman-

tic, this transformation imposes limitations on the execution order, the execution linearity and the scopes of the variables used in the operations.

5.1.2.1 Execution order

The transformation of a simple continuation is illustrated in figure 5.2 It applies on function calls similar to the abstraction (5.1). It wraps the function fn into the function fn_{due} to return a Due, as presented in section 5.1.1.2 And it relocates the continuation in a call to the method **then**. The result is similar to the abstraction (5.2). The differences are highlighted in bold font.

$$fn([arguments], continuation) \tag{5.1}$$

↓

$$fn_{\text{due}}([arguments]).\mathbf{then}(continuation) \tag{5.2}$$

Figure 5.2 – Simple transformation

The execution order is different whether *continuation* is called synchronously, or asynchronously. If fn is synchronous, it calls the *continuation* within its execution. It might execute *statements* after executing *continuation*, before returning. If fn is asynchronous, the continuation is called after the end of the current execution, after fn . The transformation erases this difference in the execution order. In both cases, the transformation relocates the execution of *continuation* after the execution of fn . For synchronous fn , the execution order changes ; the execution of *statements* at the end of fn and the continuation switch. To preserve the execution order, the function fn must be asynchronous, or execute the continuation as the last instruction.

5.1.2.2 Execution linearity

The transformation of a chain of continuations into a chain of Dues is illustrated in figure 5.3. It transforms a nested imbrication of continuations similar to the abstraction (5.3) into a flatten chain of calls encapsulating them, as abstraction (5.4).

$$\begin{array}{ccc}
\begin{array}{l}
fn1([arguments], cont1\{ \\
\quad declare\ variable \leftarrow result \\
\quad fn2([arguments], cont2\{ \\
\quad \quad print\ variable \\
\quad \}) \\
\})
\end{array}
& \rightarrow &
\begin{array}{l}
\textbf{declare variable} \\
fn1_{\textbf{due}}([arguments]) \\
\textbf{.then}(cont1\{ \\
\quad variable \leftarrow result \\
\quad \textbf{return} fn2_{\textbf{due}}([arguments]) \\
\}) \\
\textbf{.then}(cont2\{ \\
\quad print\ variable \\
\})
\end{array}
\end{array}
\tag{5.3}$$

$$\tag{5.4}$$

Figure 5.3 – Composition transformation

The main control flow characteristics, is to allow the execution order to be different from the linearity expressed in the source file. The equivalence takes into account these disruptions when modifying the source.

A call inside a loop yields multiple Dues because of the repetition, while only one can be returned to continue the chain. The others would be discarded. Similarly, a function definition is not executed in situ. It breaks the execution linearity, and prevents a call nested within it to return the Due expected to continue the chain in the parent. Therefore, the composition transformation doesn't apply on chain of Dues nested inside loops or function definitions. It must breaks the chain into simple transformation.

On the other hand, conditional branching leaves the semantic intact in a chain of Dues. If the nested asynchronous function is not called due to branching, the execution of the chain stops as expected. The transformation to a chain of Dues doesn't impact the semantic.

5.1.2.3 Variable scope

In abstraction (5.3), the definitions of *cont1* and *cont2* are overlapping. The *variable* declared in *cont1* is accessible in *cont2* to be printed. In abstraction (5.4), however, definitions of *cont1* and *cont2* are not overlapping, they are siblings. The *variable* is not accessible to *cont2*. It must be relocated in a parent function to be accessible by both *cont1* and *cont2*. The detection of such variables requires to infer their scope

statically. Most imperative languages like C/C++, Python, Ruby or Java present a lexical scope, which defines variables scopes statically. In Javascript, however, the statements `with` and `eval` modify the scope dynamically. The equivalence excludes programs using these statements to keep a lexical scope and be able to infer variable scope statically.

5.1.3 Due Compiler

The Due compiler automates the application of this equivalence on existing Javascript projects. The compilation process is made of two important steps, the identification of the continuations, and the generation of chains. The compiler is available online to reproduce the tests at compiler-due.apps.zone52.org, and the code is available online at <https://github.com/etnbrd/due-compiler>.

5.1.3.1 Identification of continuations

The first compilation step is to identify the continuations and their imbrications. The compiler transforms only *in situ* continuations - these are lambdas. Modifying continuations that are named functions break the correctness of the application. The modifications operated by the compiler modify the behavior. The modified continuations might be used elsewhere, where the modifications are not expected, hence impact the semantic.

To detect continuations, the compiler looks for callbacks. Not all detected callbacks are continuations, whereas the equivalence is applicable only on the latter. A continuation is a callback invoked only once, asynchronously. This definition is only semantical. There is no syntactical difference between a synchronous and an asynchronous callee. And it is impossible to assure a callback to be invoked only once, because the implementation of the callee is often statically unavailable.

To recognize the two, the compiler would need to have a deep understanding of the semantic of the application. Because of the highly dynamic nature of Javascript, this understanding is either unsound, limited, or complex. For example, the *node.js* method `fs.readFile` is asynchronous. Its argument is a continuation. But it can be overridden by developers to point to its synchronus counter parts. This example is very unlikely, but it shows the limitation.

The compiler identifies callback and leaves the identification of compatible continuations among these callbacks to the developer.

5.1.3.2 Generation of chains

Continuations structure the execution flow as a tree, whereas a chain of Dues arranges it sequentially. A parent continuation can execute several children, while a Due allows to chain only one. The second compilation step is to identify the trees of continuations, and trim the extra branches to transform them into chains.

If a continuation has more than one child, the compiler tries to find a single legitimate child to form the longest chain possible. A legitimate child is a unique continuation which contains another continuation to chain. If there are several continuations that continue the chain, none are the legitimate child. And the non legitimate children start new chains of Dues. In figure 5.4, the continuation **cont2** is a legitimate child, whereas **cont3** and **cont4** are not because they are siblings, and none have children. The compiler cannot decide to continue the chain with **cont3** or **cont4**, so it leaves them as is.

This step transforms each tree of continuations into several chains of continuations that translate into sequences of Dues.

```

1 caller1([args], function cont1(){
2   // ① ...
3   caller2([args], function cont2(){
4     // ③ ...
5     caller3([args], function cont3(){
6       // ⑤ ...
7     });
8     // ④ ...
9     caller4([args], function cont4(){
10      // ⑥ ...
11    });
12  });
13 // ② ...
14 })

```

Listing 5.4 - Nested calls of continuations

→

```

1 caller1([args])
2 .then(function cont1(){
3   // ① ...
4   return caller2([args])
5   // ② ...
6 })
7 .then(function cont2(){
8   // ③ ...
9   caller3([args], function cont3(){
10    // ⑤ ...
11  });
12 // ④ ...
13 caller4([args], function cont4(){
14   // ⑥ ...
15 });
16 })

```

Listing 5.5 - Chain of Due

Figure 5.4 - Transformation of a tree of continuations into a chain of Due

×

Because Dues are a placeholder for a single outcome, it doesn't allow to express the recurrence of streaming data. Concretely, a Due is created for each datum in

the stream. Therefore, continuations doesn't represent the whole streaming pipeline underlying in the application. Listeners are the callbacks starting the streaming pipeline. There is a need for an abstraction for both continuations and listeners.

Moreover, the Dues rely on a shared memory to communicate. They don't enforce the memory isolation required for parallelism.

The Due compiler is an intermediary step toward the fluxional compiler presented in section 5.2. This second compiler forms a pipeline with both listeners and continuations, and isolate the stages.

5.1.3.3 Evaluation

The compiler was evaluated against a set of Javascript projects likely to contain continuations. Because the compilation requires user interaction to detect continuations, the test set was limited to a minimum of about 50 projects to conduct the test in a reasonable time. All the projects in the set were selected from the *Node Package Manager* database to restrict the set to *Node.js* projects. They all depends on the web framework *express*, but not on the most common Promises libraries such as *Bluebird*, *Q* or *Async*. They use the test frameworks *mocha* in its default configuration. These tests are used to validate the compilation results. The test set finally contains 52 projects. This subset cannot represent the wide possibilities of Javascript, but represents a majority of common cases.

This compiler has been tested over 52 *Node.js* packages from the node package manager (npm³). 43 packages were incompatible with the compiler, 9 packages were compiled with success.

Each project passes its own tests before compilation. During the compilation, the compatible continuations were manually identified among the detected callbacks. The compilation result of each project is then tested again with its unmodified test. The compilation result should pass the tests as well as the original project. This validation assures the compiler to work as expected in most common cases.

Of the 52 projects tested, more than a half, does not contain any compatible continuations. These projects use continuations, but the compiler discards them, because they are not declared *in situ*. The other projects were rejected by the compiler because they contain **with** or **eval** statements. 9 projects compiled successfully. The compiler did not fail to compile any project of the initial test set.

Over the 9 successfully compiled projects, the compiler detected 172 callbacks.

³<https://www.npmjs.com/>

56 of them were manually identified as compatible continuations. The false positives are mainly the listeners that the web applications register to react to user requests. Listeners represent the initiation of stream of data, and are addressed in the next section.

One project contains 20 continuations, the others contains between 1 and 9 continuations each. On the 56 continuations, 36 are single. The others 20 continuations belong to imbrications of 2 to 4 continuations. The result of this evaluation prove the compiler to be able to successfully transform imbrications of continuations.

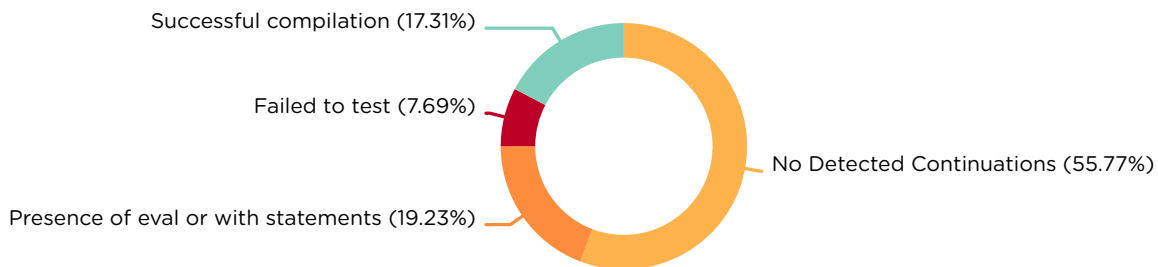


Figure 5.5 - Results of the Due compiler evaluation

On the 52 projects composing the test set

29 (55.77%) do not contain any compatible continuations,

10 (19.23%) are not compilable because they contain **with** or **eval** statements,

4 (7.69%) fail their tests before the compilation, and

9 (17.31%) compile successfully.

The compiler do not fail to compile any project. The following tables details the results of the evaluation.

29 projects contain no compatible continuation.

app-json-fetcher	release 2.1.0
brokowski	release 0.1.8
claus	release 0.0.2
code-connect-server	release 0.1.7
costa	release 0.5.0

csp-endpoint	release 0.0.1
express-device	release 0.3.11
express-resource-plus	release 0.2.1
fizzbuzz-hypermedia-server	release 0.1.2
flair-doc	release 0.1.0
generator-wikismith	release 0.2.2
gsls-transition-minify	release 0.3.1
heroku-proxy	release 2.1.1
http-test-servers	release 0.0.12
jellyjs-plugin-httpserver	release 0.0.8
loopback-angular-cli	release 1.1.1
loopback-explorer	release 1.6.4
moby	release 1.1.1
monami	release 0.0.21
mongoose-express-resource	release 0.1.2
nodebootstrap-server	release 1.1.2
oauth-express	release 0.0.1
public-server	release 2.0.1
scrapit	release 0.0.4
sik	release 0.0.1
sonea	release 0.0.3
squirrel-server	release 0.0.1
vsoft-explorer	release 0.2.1
webs-weeia	release 0.2.2

10 projects contain eval or with statements.

adnoce	release 1.1.2
arkhaio5	release 0.2.0
browserman	release 0.1.2
infectwit	release 0.0.1
ldapp	release 0.1.4
levelhud	release 0.1.3
manet	release 0.3.2
solid	release 0.2.1
swac	release 0.12.1
swac-odm	release 0.2.1

4 projects fail their tests before the compilation.

express-orm-mvc	release 1.1.0
derpjs	release 0.2.2
hangout	release 0.0.3
ord.zeke.xxx	release 1.0.0

9 projects successfully compile.

express-user-couchdb	release 0.3.5
express-endpoint	release 1.2.11
gifsockets-server	release 0.38.1
heroku-bouncer	release 4.0.1
moonridge	release 0.6.9
redis-key-overview	release 0.0.3
slack-integrator	release 0.0.6
timbits	release 0.7.3

The compiler detected 172 callbacks, 52 of them turned out to be compatible continuations.

name	continuations		chains length			
	total	compiled	1	2	3	4
express-user-couchdb	40	20	9	2	1	1
express-endpoint	19	2	2			
gifsockets-server	3	1	1			
heroku-bouncer	7	3	3			
moonridge	37	6	2	2		
redis-key-overview	14	9	6		1	
slack-integrator	6	3	1	1		
timbits	34	8	8			
tingo-rest	12	4	4			
total	172	54	36	5	2	1

In the following tables, are the name of the callee for each continuations, grouped by package.

express-endpoint	ruleFn
	parseParams
	config.validateUser
	createSession
	db.destroy
express-user-couchdb	db.get ×5
	db.insert ×5
	db.view ×3

	getUserName
	lookupUser
	req.session.destroy ×2
gifsockets-server	getRawBody
	ensureValidToken
heroky-bouncer	oauth.getOAuthAccessToken
	request.post
	doc.remove
	doc.save
moonridge	model.findById ×2
	mongoose.connect
	populateWithClientQuery
	exec ×2
	execstderr
redis-key-overview	fs.unlink
	fs.writeFile ×2
	request.post
	res.sendFile ×3
	config.payload
slack-integrator	sendPayload
	request
	loadTimbits
	res.render
timbits	timbit.test ×2
	timbits.pantry.fetch ×2
	request
	req.collection.findOne

```
req.collection.insert  
req.collection.update  
req.collection.remove
```

5.2 Fluxions

The second contribution of this thesis is the equivalence between a global memory abstraction and a distributed memory. It tackles the problems arising from the replacement of the global memory synchronizations with message passing.

This equivalence is implemented as a compiler, improving upon the previous one. The compiler transforms a Javascript application into a network of independent parts communicating by message streams and executed in parallel.

Like the previous compiler, this compiler is organized into two main steps. The first step is the identification of the rupture points between fluxions, addressed in section 5.2.1. The second step is the isolation between the fluxions, addressed in section 5.2.2. The compiler is tested on a real-case, to expose its limits in section 5.2.3.

5.2.1 Fluxions Compiler

The source language for this transformation is of higher-order to allow the modularity required for productivity. Moreover, it is implemented as an event-loop to imposes the developer to define the causality between asynchronous operations. The compiler transforms a *Node.js* application into a fluxional application compliant with the execution model described in section 4.2.2. The chain of compilation is described in figure 5.6.

The compiler uses the *estools*⁴ suite to parse (**esprima**), analyze (**escope**), manipulate (**estrace** and **esquery**) and generate (**escodegen**) source code from an Abstract Syntax Tree (AST). It is tailored for – but not limited to – web applications using *Express*⁵, the most used *Node.js* web framework. The compiler extracts

⁴<https://github.com/estools>

⁵<http://expressjs.com/>

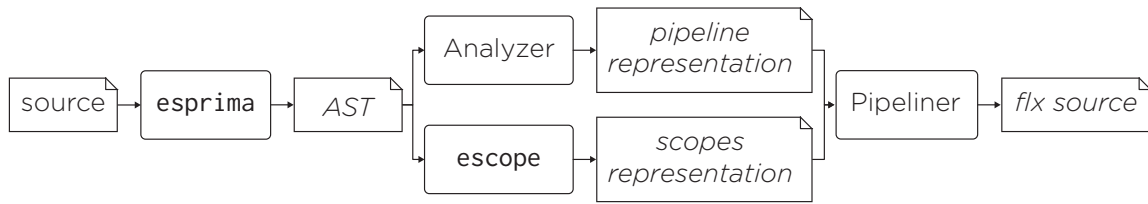


Figure 5.6 – Compilation chain

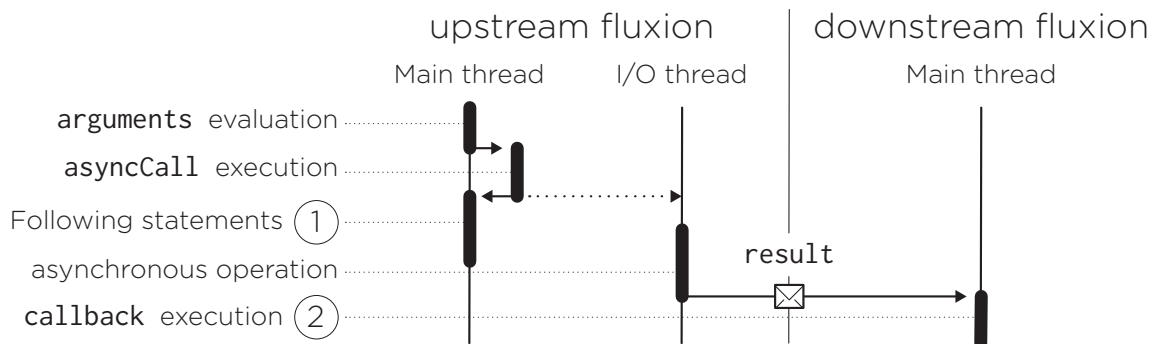
an AST from the source with **esprima**. From this AST, the *Analyzer* step identifies the rupture points the different application parts. This first step outputs a pipeline representation of the application. In this pipeline representation, the stages are not yet independent and encapsulated into fluxions. From the AST, **escope** produces a representation of the memory scopes. The *Pipeliner* step, explained in section 5.2.2, analyzes the pipeline representation and the scopes representation to distribute the shared memory into independent groups of fluxions.

5.2.1.1 Detection

In *Node.js*, I/O operations are asynchronous functions and indicate rupture points between two application parts. Figure 5.7 shows a code example of a rupture point with the illustration of the execution of the two application parts isolated into fluxions. The two application parts are the caller of the asynchronous function call on one hand, and the callback provided to the asynchronous function call on the other hand.

Similarly as in the Due compiler, the detection of asynchronous callees is done by the developer. It uses a list of common asynchronous callees, like the **express** and file system methods. This list can be augmented to match asynchronous callees individually for any application. To identify the callee, the analyzer walks the AST to find a call expression matching this list.

After the identification of the callee, the callback needs to be identified as well, to be encapsulated in the downstream fluxion. For each asynchronous call detected, the compiler tests if one of the arguments is of type **function**. The callback functions declared *in situ* are trivially detected in the AST. The compiler discard callbacks not declared *in situ*, to avoid altering the semantic by moving or modifying their definitions.



```

1 asyncCall(arguments, function callback(result){ ② });
2 // Following statements ①

```

Figure 5.7 – Rupture point interface

5.2.2 Fluxions Isolation

As a rupture point occur between an asynchronous caller and a callback defined *in situ*, it eventually breaks the chain of scopes. If the caller and the callback are separated, it breaks the closures of the callback. The callback in the downstream fluxion cannot access the scope of its parent as expected. The pipeliner step replaces the need for this closure, allowing application parts to be isolated, and to rely only on independent memory stores and message passing. It determines the distribution using the scope representation, which represents the variables' dependencies between application parts. Depending on this representation, the compiler can replace the broken closures in three different ways. We present these three alternatives in figure 5.8.

5.2.2.1 Scope

If a variable is modified inside only one application part in the current *post* chain, then the pipeliner adds it to the context of its fluxion.

In figure 5.8, the variable `a` is updated in the function `add`. The pipeliner step stores this variable in the context of the fluxion `add`.

5.2.2.2 Stream

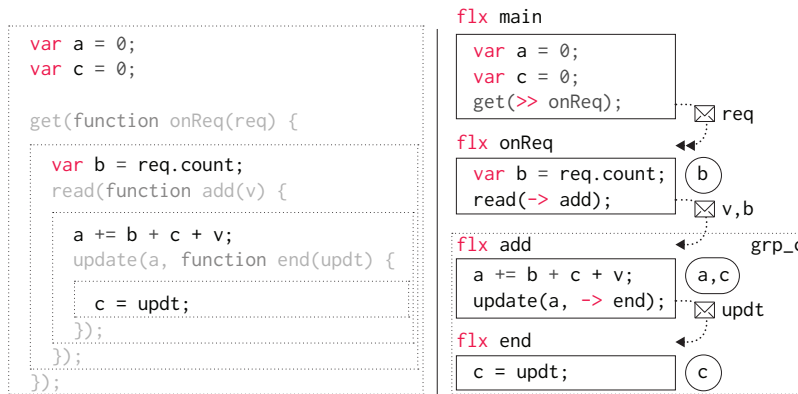


Figure 5.8 - Variable management from Javascript to the high-level fluxional language

If a modified variable is read by some downstream application parts, then the pipeliner makes the upstream fluxion add this variable to the message stream to be sent to the downstream fluxions. It is impossible to send variables to upstream fluxions, without causing inconsistencies. If the fluxion retro propagates the variable for an upstream fluxion to read, the upstream fluxion might use the old version while the new version is on its way.

In figure 5.8, the variable **b** is set in the function **onReq**, and read in the function **add**. The pipeliner step makes the fluxion **onReq** send the updated variable **b**, in addition to the variable **v**, in the message sent to the fluxion **add**.

Exceptionally, if a variable is defined inside a *post* chain, like **b**, then this variable can be streamed inside this *post* chain without restriction on the order of modification and read. Indeed, in the current *post* chain, the execution of the upstream fluxion is assured to end before the execution of the downstream fluxion, because of their causality. Therefore, no reading of the variable by the upstream fluxion happens after the modification by the downstream fluxion.

5.2.2.3 Share

If a variable is needed for modification by several application parts, or is read by an upstream application part, then it needs to be synchronized between the fluxions. The pipeliner groups all the fluxions sharing this variable with the same tag. And it adds this variable to the contexts of each fluxions.

In figure 5.8, the variable **c** is set in the function **end**, and read in the function **add**. As the fluxion **add** is upstream of **end**, the pipeliner step groups the fluxion **add**

and end with the tag `grp_c` to allow the two fluxions to share this variable.

5.2.3 Real test case

The compiler is tested on a real application, `gifsockets-server`⁶. This test proves the possibility for an application to be compiled into a network of independent parts. It shows the current limitations of this isolation and the modifications needed on the application to circumvent them.

```
1 var express = require('express'),
2   app = express(),
3   routes = require('gifsockets-middleware'),
4   getRawBody = require('raw-body');
5
6 function bodyParser(limit) {
7   return function saveBody(req, res, next) {
8     getRawBody(req, {
9       expected: req.headers['content-length'],
10      limit: limit
11    }, function (err, buffer) {
12      req.body = buffer;
13      next();
14    });
15  };
16 }
17
18 app.post('/image/text', bodyParser(1 * 1024 * 1024), routes.writeTextToImages);
19 app.listen(8000);
```

Listing 5.6 – Simplified version of `gifsockets-server`

This application, simplified in listing 5.6, is a real-time chat using gif-based communication channels. It was selected from the evaluation set of the Due compiler because it is simple enough to illustrate this evaluation. The server transforms the received text into a gif frame, and pushes it back to a never-ending gif to be displayed on the client.

On line 18, the application registers two functions to process the requests received on the url `/image/text`. The closure `saveBody`, line 7, returned by `bodyParser`, line 6, and the method `routes.writeTextToImages` from the external module `gifsockets-middleware`, line 3. The closure `saveBody` calls the asynchronous function `getRawBody` to get the request body. Its callback handles the errors, and calls `next` to continue processing the request with the next function, `routes.writeTextToImages`.

⁶<https://github.com/twolfson/gifsockets-server>

5.2.3.1 Compilation

The compilation result is in listing 5.7. The function call `app.post`, line 18, is a rupture point. However, its callbacks, `bodyParser` and `routes.writeTextToImages` are not declared *in situ*. They are evaluated as functions only at runtime. As precised previously, the compiler discards these callbacks to avoid altering the semantic.

```
1 flx main & express {req}
2 » anonymous_1000 [req, next]
3   var express = require('express'),
4     app = express(),
5     routes = require('gifsockets-middleware'), //
6     getRawBody = require('raw-body');
7
8   function bodyParser(limit) { //
9     return function saveBody(req, res, next) { //
10       getRawBody(req, { //
11         expected: req.headers['content-length'], //
12         limit: limit
13       }, » anonymous_1000 [req, next]);
14     };
15   }
16
17   app.post('/image/text', bodyParser(1 * 1024 * 1024), routes.writeTextToImages); //
18   app.listen(8000);
19
20 flx anonymous_1000
21 -> null
22   function (err, buffer) { //
23     req.body = buffer; //
24     next(); //
25   }
```

Listing 5.7 – Compilation result of gifsockets-server

The compiler detects a rupture point : the function `getRawBody` and its anonymous callback, line 11. It encapsulates this callback in a fluxion named `anonymous_1000`. The callback is replaced with a stream placeholder to send the message stream to this downstream fluxion. The variables `req` and `next` are appended to this message stream, to propagate their value from the `main` fluxion to the `anonymous_1000` fluxion.

When `anonymous_1000` is not isolated from the `main` fluxion, as if they belong to the same group, the compilation result works as expected. The variables used in the fluxion, `req` and `next`, are still shared between the two fluxions. In this situation fluxions are quite similar to Dues regarding memory shareing. Our goal is to isolate the two fluxions, to be able to safely parallelize their executions.

5.2.3.2 Isolation

In listing 5.7, the fluxion `anonymous_1000` modifies the object `req`, line 23, to store the text of the received request, and it calls `next` to continue the execution, line 24. `req` is an alias to a memory location used in multiple places in code. Therefore, these operations produce side-effects that should propagate in the whole application, but the isolation prevents this propagation. Isolating the fluxion `anonymous_1000` produces runtime exceptions. The next paragraph details how this situation is handled to allow the application to be parallelized.

Variable req

The variable `req` is read in fluxion `main`, lines 10 and 11. Then its property `body` is associated to `buffer` in fluxion `anonymous_1000`, line 23. The compiler is unable to identify the aliases of this variable. However, the side effect resulting from this association impacts a variable in the scope of the next callback, `routes.writeTextToImages`. In this text case, the application is modified manually to explicitly propagate this side-effect to the next callback through the function `next`. The modifications of this function are explained further in the next paragraph.

Closure next

The function `next` is a closure provided by the `express Router` to continue the execution with the next function to handle the client request. Because it indirectly relies on the variable `req`, it is impossible to isolate its execution with the `anonymous_1000` fluxion. Instead, we modify `express`, so as to be compatible with the fluxional execution model. We explain the modifications below.

```
1 flx anonymous_1000
2 -> express_dispatcher
3   function (err, buffer) { //
4     req.body = buffer; //
5     next_placeholder(req, -> express_dispatcher); //
6   }
7
8 flx express_dispatcher & express {req} //
9 -> null
10  function (modified_req) {
11    merge(req, modified_req);
12    next(); //
13  }
```

Listing 5.8 – Simplified modification on the compiled result

In listing 5.6, the function `next` is a continuation allowing the anonymous callback, line 11, to call the next function to handle the request. To isolate the anonymous callback into `anonymous_1000`, `next` is replaced by a rupture point. This replacement is illustrated in listing 5.8. The `express Router` registers a fluxion named `express_dispatcher`, line 8, to continue the execution after the fluxion `anonymous_1000`. This fluxion is in the same group `express` as the `main` fluxion, hence it has access to the original variable `req`, and to the original function `next`. The call to the original `next` function is replaced by a placeholder to push the stream to the fluxion `express_dispatcher`, line 5. The fluxion `express_dispatcher` receives the stream from the upstream fluxion `anonymous_1000`, merges back the modification in the variable `req` to propagate the side effects, and finally calls the original function `next` to continue the execution, line 12.

After the modifications detailed above, the server works as expected. The isolated fluxion correctly receives, and returns its serialized messages. The client successfully receives a gif frame containing the text.

5.3 Evaluation

The solution previously presented is evaluated against the criteria presented in chapter 3. The criteria are Productivity, Efficiency and Adoption.

5.3.1 Productivity and Adoption

The solution presented intends to disrupt as less as possible the way developer build web applications. The goal is to avoid degrading the productivity, hence the adoption, of the proposed platform.

In the current state, the compiler implementation is unable to operate the transformation without the help of the developer. This limitation is due to the static analysis unable to correctly detect the aliasing of the memory in Javascript.

However, with a dynamic analysis at runtime, this detection should improve. And it would allow to operate the transformation without the help of the developer.

In this case, because the productivity is left untouched, this platform should be able to leverage the adoption of Javascript without trouble. Moreover, Worldline could be able to propose a service based on this transformation. A scalable PaaS,

without the need for the developer to write stateless applications, or other current inconvenient.

Model	Composition	Encapsulation	→	Productivity
Fluxional Compiler	2	5		2
Fluxional Runtime	5	5		5

Table 5.7 – Productivity of the proposed solution

5.3.2 Efficiency

The implementation of the compiler is not finished enough to compile a real application without help from the developer. It doesn't make any sense to evaluate an application, as the transformation would not reflect the compilation process, but the manual transformation process. Indeed, it is already known that distributed application can have very good performance efficiency.

If the runtime memory analysis is solid enough to detect correctly the aliasing of the memory, then it will be able to help the development team transitioning from productivity to efficiency, which is the response of this thesis to the problematic.

Model	Fine-grain level synchronization	Coarse-grain level message passing	→	Efficiency
Fluxional Compiler	5	3		3
Fluxional Runtime	5	5		5

Table 5.8 – Efficiency of the proposed solution

5.3.3 Summary

This thesis was unfortunately unable to completely address the problematic in time, and leave the science in an unbearable doubt. Is it possible to have a smooth transition from productivity to efficiency. The author dares to say yes, but the humanity is not ready yet.

Model	Productivity	Efficiency	Adoption
Fluxional Compiler	2	3	3
Fluxional Runtime	5	5	3

Table 5.9 – Sommary of the proposed solution

5.3.4 Perspectives

Javascript is a highly dynamic languages. Chapter 3 presented that the late binding provided by this dynamism is an advantage for the modularity, hence for productivity. However, this late binding makes the static analysis eventually impossible.

5.3.5 Limitations

5.3.5.1 Levels of Indirections

The indirection is an abstraction between the value, and its manipulation. In listing 6.1, the variables **a** and **b** point both to the same memory object. The function **fn** introduces a level of indirection between the real object **a** and its manipulation handle, **b**; Actually, the variable **a** already introduces a level of indirection between the real object and the handle **a**.

```

1 var a = {
2     // an object;
3 };
4
5 fn(b) {
6     // modify b;
7 }
8
9 fn(a);

```

Listing 5.9 – One level of Indirection

5.3.5.2 Incertitudes

The level of indirection is trivial in listing 6.1. A very simple static analysis could resolve it. It only needs to have access to the definition of **a** and of **fn**. However, in listing 6.2, the array **handlers** introduces a new level of indirection. Additionally to **a**, and **fn** in **handlers**, the static analysis now needs to have access to the definition of **i**. If this definition is provided by an external input, it is not available statically, hence, it adds an incertitude during the analysis.

```

1 var a = {
2     // an object;
3 },
4     handlers = [
5     // definition of fn handlers;
6 ],
7     i = ?;
8
9 handlers[i](a);
10 handlers[i+1](a);

```

Listing 5.10 – Two levels of indirection

These examples are extremely simplified. In practice, a program contains enough indirections, for the static analysis to be overwhelmed by incertitudes. These incertitudes limit the fluxional compiler presented in chapter ???. It is unable to detect correctly most of the rupture points, not to discover the interdependencies between fluxions.

If one of the **handlers** is an asynchronous function that could yield a rupture point. The detection of this rupture point requires no incertitude on the variable

i. In a web application, this variable `i` is probably part of the user request, which eventually introduces an incertitude. Moreover, if this incertitudes could be resolved somehow, the isolation of the downstream fluxion would require no incertitude on the variables shared with the scope of upstream fluxion. This is very unlikely.

No static analysis can really overstep these limitations. Only a dynamic analysis could analysis the resolved indirections during run time to overstep these limitations correctly.

5.3.6 Dynamic Analysis and Compilation _____

Javascript is compiled at run time to improve performances. During this compilation, all the levels of indirections are resolved to actually manipulate the memory. As future works, we intend to implement the fluxional compiler as parts of a just-in-time compilation in a Javascript execution engine.

5.3.6.1 Just-in-time Compilation

With the dynamic registering of Fluxions to the messaging system, and into tag groups, it is possible to transform a Javascript application continuously during its execution. Analysis of the interdependencies become as trivial as for static languages, with the resolution of the indirections by the just-in-time compiler. The fluxional compiler waits for these resolutions, and then analyzes the compiled code for rupture points. As the asynchronism of a function call is handled by the execution engine, the just-in-time compilation can pin point precisely the asynchronous calls from the synchronous ones. And the continuations for these asynchronous calls are resolved, which makes them similar to inline continuations.

Moreover, this memory analysis would allow the closure serialization required for the use of higher-order functions in the fluxional execution model.

5.3.6.2 Dynamic Grouping

The isolation of a fluxion can also be done during the execution. If it depends on an upstream variable, the fluxion is dynamically registered to the group, and can be deployed.

If the new fluxion depends on a local variable, as well as a variable from a group on another node than the local node, the group needs to be deployed back locally. The fluxion as well as all the fluxions of the group are deployed locally, but the execution needs to wait for the contexts of the group to be available locally. To gather the contexts, the node responsible for this group send a message to the messaging system managing this group. The messaging system gather all the contexts of the fluxions, and send them back. When the contexts are deployed locally to the new node responsible for the this group, the execution of this branch can resume.

A new compromise have to be done between the cost of sending a fluxion and the cost to get it back, and the risk that it gets sent back. It might be possible to reduce this risk by saving the compilation information from one execution to the other.

CHAPTER 6

CONCLUSION

Contents

6.1 Summary	111
6.1.1 Fluxional Execution Model	111
6.1.2 Pipeline Extraction	112
6.1.3 Pipeline Isolation	112
6.2 Perspectives	112
6.3 Perspectives	112
6.3.1 Limitations	113
6.3.1.1 Levels of Indirections	113
6.3.1.2 Incertitudes	113
6.3.2 Dynamic Analysis and Compilation	114
6.3.2.1 Just-in-time Compilation	114
6.3.2.2 Dynamic Grouping	115
6.4 Opening	115

The web brought a new economic model allowing a tremendous number of business opportunities. To seize these opportunities, a team needs to develop a web application, and grow a business around it. The economical incentives around the technical development changes completely during the growth of this business. In the beginning, the development needs to be productive, to quickly release a product and iterate with the user feedbacks. When the project matures, the execution needs to be efficient, to cope with the load of a large user base while limiting the hardware costs.

These two development concerns are incompatible. No platform can provide both performance efficiency, and development productivity at the same time. The platforms at the state of the art propose only a fixed compromise between the two. This thesis presented a platform allowing an evolutive compromise to fit the economical incentives.

6.1 Summary

The main contribution of this thesis is the equivalence to allow a continuous shift from productivity to efficiency. The equivalence map programs written for the event-driven execution model onto the pipeline execution model. It is based on the identification of rupture point in the synchronous control-flow. These ruptures points are the boundaries between the stages of the underlying pipeline. The equivalence extracts, and isolates these stages to execute them in parallel in the fluxional execution model.

6.1.1 Fluxional Execution Model

The fluxional execution model executes an application expressed as a network of fluxions. That is an independent execution container at the function level, communicating by message. The fluxions are executed in parallel to distribute the computation across several cores and increase the performance efficiency of the application.

6.1.2 Pipeline Extraction

The compiler identifies and extracts the stages of the asynchronous control flow. The first step of this compiler identifies these stages to express them as a pipeline.

The difficulty in this compilation step is to identify the asynchronous functions indicating the stages. Because of the dynamic nature of Javascript, it is impossible to statically detect these functions. The compiler is unable to reliably detect them. Instead the compiler relies on the developer to provide a list of asynchronous function names to extract.

6.1.3 Pipeline Isolation

The compiler then isolates each stage so that it can be executed in parallel, on the fluxional execution model.

The execution flows from one stage to the other, from the reception of a request until the last stage responds to the request. The stages are isolated, and communicate only by message along this pipeline. To communicate the request, and the other state updates, they must follow some rules. If a stage needs to access variable from one request to the other, this variable is stored in its context. If a downstream stage needs to read a variable from an upstream stage, the variable is sent as a message in the data-flow of the application. If two stages need to share a variable, they are grouped on the same execution node to safely share parts of their context. Their executions are not parallelized to avoid conflicting accesses.

The difficulty in this step is to identify the memory dependencies between stages. Because of the dynamic nature of Javascript, there is a lot of possible aliasing in the memory. The compiler, in its current state is unable to overcome this difficulty.

6.2 Economic Considerations

This thesis intends to bring a reconciliation between two economical concerns in the development of a web application, the efficiency of execution and the productivity of development. A third concern is currently increasingly taking importance. The IT industry has an increasing carbon footprint which has an important impact on the environment. As the IT industry permeates into literally every aspect of our lives, it will be of crucial importance to consider this impact for future developments.

BIBLIOGRAPHY

- [1] Sebastian Adam and Joerg Doerr. “How to better align BPM & SOA - Ideas on improving the transition between process design and deployment”. In: *CEUR Workshop Proceedings*. Vol. 335. 2008, pp. 49–55.
- [2] A Adya, J Howell, and M Theimer. “Cooperative Task Management Without Manual Stack Management.” In: *USENIX Annual Technical Conference* (2002).
- [3] Yuichiro Ajima, Takafumi Nose, Kazushige Saga, Naoyuki Shida, and Shinji Sumimoto. “ACPdI”. In: *Proceedings of the First International Workshop on Extreme Scale Programming Models and Middleware - ESPM '15*. New York, New York, USA: ACM Press, Nov. 2015, pp. 11–18. DOI: 10.1145/2832241.2832242.
- [4] T Akidau and A Balikov. “MillWheel: Fault-Tolerant Stream Processing at Internet Scale”. In: *Proceedings of the VLDB Endowment 6.11* (2013).
- [5] Frances E. Allen. “Control flow analysis”. In: *ACM SIGPLAN Notices* 5.7 (July 1970), pp. 1–19. DOI: 10.1145/390013.808479.
- [6] SP Amarasinghe, JAM Anderson, MS Lam, and CW Tseng. “An Overview of the SUIF Compiler for Scalable Parallel Machines.” In: *PPSC* (1995).
- [7] Gene M. Amdahl. “Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities”. In: *AFIPS Spring Joint Computer Conference, 1967. AFIPS '67 (Spring). Proceedings of the*. Vol. 30. 1967, pp. 483–485. DOI: doi:10.1145/1465482.1465560.
- [8] LO Andersen. “Program analysis and specialization for the C programming language”. In: (1994).

- [9] James H. Anderson and Mohamed G. Gouda. *The virtue of Patience: Concurrent Programming With And Without Waiting*. 1990.
- [10] Joe Armstrong. *Programming Erlang*. Pragmatic Programmers, 2007, p. 519. DOI: [10.1017/S0956796809007163](https://doi.org/10.1017/S0956796809007163).
- [11] Joe Armstrong, Robert Virding, Claes Wikstrom, and Mike Williams. *Concurrent Programming in ERLANG*. 1993.
- [12] Michel Auguin and Francois Larbey. “OPSILA: an advanced SIMD for numerical analysis and signal processing”. In: *Microcomputers: developments in industry, business, and education*. 1983, pp. 311–318.
- [13] U Banerjee. *Loop parallelization*. 2013.
- [14] Cédric Bastoul, Albert Cohen, Sylvain Girbal, Saurabh Sharma, and Olivier Temam. “Putting Polyhedral Loop Transformations to Work”. In: *LCPC '04 Languages and Compilers for Parallel Computing*. Lecture Notes in Computer Science 2958. Chapter 14 (2004). Ed. by Lawrence Rauchwerger, pp. 209–225. DOI: [10.1007/b95707](https://doi.org/10.1007/b95707).
- [15] Michael Bauer, Sean Treichler, Elliott Slaughter, and Alex Aiken. “Legion: Expressing Locality and Independence with Logical Regions”. In: *Proceedings of the International Conference on High Performance Computing Networking Storage and Analysis SC 12* (Nov. 2012), pp. 1–11. DOI: [10.1109/SC.2012.71](https://doi.org/10.1109/SC.2012.71).
- [16] Micah Beck, Richard Johnson, and Keshav Pingali. “From control flow to dataflow”. In: *Journal of Parallel and Distributed Computing* 12.2 (1991), pp. 118–129. DOI: [10.1016/0743-7315\(91\)90016-3](https://doi.org/10.1016/0743-7315(91)90016-3).
- [17] JR von Behren, J Condit, and EA Brewer. “Why Events Are a Bad Idea (for High-Concurrency Servers).” In: *HotOS* (2003).
- [18] R Von Behren, J Condit, and F Zhou. “Capriccio: scalable threads for internet services”. In: *ACM SIGOPS ...* (2003).
- [19] M Bodin and A Charguéraud. “A trusted mechanised JavaScript specification”. In: *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (2014).
- [20] S. D. Brookes, C. A. R. Hoare, and A. W. Roscoe. “A Theory of Communicating Sequential Processes”. In: *Journal of the ACM* 31.3 (June 1984), pp. 560–599. DOI: [10.1145/828.833](https://doi.org/10.1145/828.833).
- [21] T Buddhika and S Pallickara. “NEPTUNE: Real Time Stream Processing for Internet of Things and Sensing Environments”. In: *granules.cs.colostate.edu* ().

- [22] Marcelo Cataldo, Patrick A. Wagstrom, James D. Herbsleb, and Kathleen M. Carley. “Identification of coordination requirements”. In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06*. New York, New York, USA: ACM Press, Nov. 2006, p. 353. DOI: 10.1145/1180875.1180929.
- [23] Bryan Catanzaro, Shoaib Kamil, and Yunsup Lee. “SEJITS: Getting productivity and performance with selective embedded JIT specialization”. In: ... *Models for Emerging ...* (2009), pp. 1–10. DOI: 10.1.1.212.6088.
- [24] B.L. Chamberlain, D. Callahan, and H.P. Zima. “Parallel Programmability and the Chapel Language”. In: *International Journal of High Performance Computing Applications* 21.3 (Aug. 2007), pp. 291–312. DOI: 10.1177/1094342007078442.
- [25] F Chan, J N Cao, A T S Chan, and M Y Guo. “Programming support for MPMD parallel computing in ClusterGOP”. In: *IEICE Transactions on Information and Systems* E87D.7 (2004), pp. 1693–1702.
- [26] K. Mani Chandy and Carl Kesselman. “Compositional C++: Compositional parallel programming”. In: *Languages and Compilers for Parallel Computing*. Vol. 757. 2005, pp. 124–144. DOI: 10.1007/3-540-48319-5.
- [27] Chi-Chao Chang, G. Czajkowski, T. Von Eicken, and C. Kesselman. “Evaluating the Performance Limitations of MPMD Communication”. In: *ACM/IEEE SC 1997 Conference (SC'97)* (1997), pp. 1–10. DOI: 10.1109/SC.1997.10040.
- [28] Barbara Chapman, Tony Curtis, Swaroop Pophale, Stephen Poole, Jeff Kuehn, Chuck Koelbel, and Lauren Smith. “Introducing OpenSHMEM”. In: *Proceedings of the Fourth Conference on Partitioned Global Address Space Programming Model - PGAS '10*. New York, New York, USA: ACM Press, Oct. 2010, pp. 1–3. DOI: 10.1145/2020373.2020375.
- [29] Philippe Charles, Christian Grothoff, Vijay Saraswat, Christopher Donawa, Allan Kielstra, Kemal Ebcioglu, Christoph von Praun, and Vivek Sarkar. “X10”. In: *Proceedings of the 20th annual ACM SIGPLAN conference on Object oriented programming systems languages and applications - OOPSLA '05*. Vol. 40. 10. New York, New York, USA: ACM Press, Oct. 2005, p. 519. DOI: 10.1145/1094811.1094852.
- [30] Chun Chen, Jacqueline Chame, and Mary Hall. “CHiLL: A framework for composing high-level loop transformations”. In: *U. of Southern California, Tech. Rep* (2008), pp. 1–28. DOI: 10.1001/archneur.64.6.785.

- [31] Andrey Chudnov and David A. Naumann. “Inlined Information Flow Monitoring for JavaScript”. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Oct. 2015), pp. 629–643. DOI: 10.1145/2810103.2813684.
- [32] Austin T. Clements, M. Frans Kaashoek, Nickolai Zeldovich, Robert T. Morris, and Eddie Kohler. “The scalable commutativity rule”. In: *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles - SOSOP ’13*. New York, New York, USA: ACM Press, Nov. 2013, pp. 1–17. DOI: 10.1145/2517349.2522712.
- [33] William Douglas Clinger. “Foundations of Actor Semantics”. eng. In: (May 1981).
- [34] M. I. Cole. *Algorithmic skeletons : A structured approach to the management of parallel computation*. eng. 1988.
- [35] C Consel, H Hamdi, and L Réveillère. “Spidle: a DSL approach to specifying streaming applications”. In: *Generative ...* (2003).
- [36] Melvin E. Conway. “Design of a separable transition-diagram compiler”. In: *Communications of the ACM* 6.7 (July 1963), pp. 396–408. DOI: 10.1145/366663.366704.
- [37] David E. Culler, A. Dusseau, Seth Copen Goldstein, Arvind Krishnamurthy, Steven Lumetta, Thorsten Von Eicken, and Katherine Yelick. “Parallel programming in Split-C”. English. In: (), pp. 262–273. DOI: 10.1109/SUPERC.1993.1263470.
- [38] Frank Dabek and Nickolai Zeldovich. “Event-driven programming for robust software”. In: *Proceedings of the 10th workshop on ACM SIGOPS European workshopn workshop* (July 2002), pp. 186–189. DOI: 10.1145/1133373.1133410.
- [39] F. Darema, D.A. George, V.A. Norton, and G.F. Pfister. “A single-program-multiple-data computational model for EPEX/FORTRAN”. In: *Parallel Computing* 7.1 (Apr. 1988), pp. 11–24. DOI: 10.1016/0167-8191(88)90094-4.
- [40] Frederica Darema. “The SPMD Model: Past , Present and Future”. In: *Parallel Computing*. 2001, p. 1. DOI: 10.1007/3-540-45417-9{_}1.
- [41] Jeffrey Dean and Sanjay Ghemawat. “MapReduce: Simplified Data Processing on Large Clusters”. In: *Proc. of the OSDI - Symp. on Operating Systems Design and Implementation*. Vol. 51. 1. 2004, pp. 137–149. DOI: 10.1145/1327452.1327492. arXiv: 10.1.1.163.5292.

- [42] E W Dijkstra. *Notes on structured programming*. 1970.
- [43] Edsger Dijkstra. “Over de sequentialiteit van procesbeschrijvingen”. In: ().
- [44] Edsger W. Dijkstra. “Guarded commands, nondeterminacy and formal derivation of programs”. In: *Communications of the ACM* 18.8 (Aug. 1975), pp. 453–457. DOI: 10.1145/360933.360975.
- [45] Edsger W. Dijkstra. “Letters to the editor: go to statement considered harmful”. In: *Communications of the ACM* 11.3 (Mar. 1968), pp. 147–148. DOI: 10.1145/362929.362947.
- [46] Edsger W. Dijkstra. “The structure of the “THE”-multiprogramming system”. In: *Communications of the ACM* 11.5 (May 1968), pp. 341–346. DOI: 10.1145/363095.363143.
- [47] Julian Dolby. “A History of JavaScript Static Analysis with WALA at IBM”. In: (2015).
- [48] H. Carter Edwards and Daniel Sunderland. “Kokkos Array performance-portable manycore programming model”. In: *Proceedings of the 2012 International Workshop on Programming Models and Applications for Multicores and Manycores - PMAM '12*. New York, New York, USA: ACM Press, Feb. 2012, pp. 1–10. DOI: 10.1145/2141702.2141703.
- [49] Raul Castro Fernandez, Matteo Migliavacca, Evangelia Kalyvianaki, and Peter Pietzuch. “Making state explicit for imperative big data processing”. In: *USENIX ATC* (2014).
- [50] JI Fernández-Villamor. “Microservices-Lightweight Service Descriptions for REST Architectural Style.” In: ... *2010-Proceedings of ...* (2010).
- [51] D Flanagan. *JavaScript: the definitive guide*. 2006.
- [52] Michael J. Flynn. “Some Computer Organizations and Their Effectiveness”. English. In: *IEEE Transactions on Computers* C-21.9 (Sept. 1972), pp. 948–960. DOI: 10.1109/TC.1972.5009071.
- [53] Ian Foster, Carl Kesselman, and Steven Tuecke. “The Nexus Approach to Integrating Multithreading and Communication”. In: *Journal of Parallel and Distributed Computing* 37.1 (Aug. 1996), pp. 70–82. DOI: 10.1006/jpdc.1996.0108.
- [54] I.T. Foster and K M Chandy. “Fortran M: A Language for Modular Parallel Programming”. In: *Journal of Parallel and Distributed Computing* 26.1 (Apr. 1995), pp. 24–35. DOI: 10.1006/jpdc.1995.1044.

- [55] M Fowler and J Lewis. “Microservices”. In: <http://martinfowler.com/articles/microservices.html> [... (2014).
- [56] Matteo Frigo, Charles E. Leiserson, and Keith H. Randall. “The implementation of the Cilk-5 multithreaded language”. In: *ACM SIGPLAN Notices* 33.5 (May 1998), pp. 212–223. DOI: 10.1145/277652.277725. arXiv: 9809069v1 [arXiv:gr-qc].
- [57] P Gardner and G Smith. “JuS: Squeezing the sense out of javascript programs”. In: *JSTools@ ECOOP* (2013).
- [58] PA Gardner, S Maffeis, and GD Smith. “Towards a program logic for JavaScript”. In: *ACM SIGPLAN Notices* (2012).
- [59] JJ Garrett. “Ajax: A new approach to web applications”. In: (2005).
- [60] Tarek El-Ghazawi and Lauren Smith. “UPC: unified parallel C”. In: *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06*. New York, New York, USA: ACM Press, Nov. 2006, p. 27. DOI: 10.1145/1188455.1188483.
- [61] Adele Goldberg. *Smalltalk-80 : the interactive programming environment*. 1984, xi, 516 p.
- [62] Horacio González-Vélez and Mario Leyton. “A survey of algorithmic skeleton frameworks: high-level structured parallel programming enablers”. In: *Software: Practice and Experience* 40.12 (Nov. 2010), pp. 1135–1160. DOI: 10.1002/spe.1026.
- [63] J Gosling. *The Java language specification*. 2000.
- [64] Steven D. Gribble, Matt Welsh, Rob Von Behren, Eric a. Brewer, David Culler, N. Borisov, S. Czerwinski, R. Gummadi, J. Hill, A. Joseph, R. H. Katz, Z. M. Mao, S. Ross, and B. Zhao. “Ninja architecture for robust Internet-scale systems and services”. In: *Computer Networks* 35.4 (2001), pp. 473–497. DOI: 10.1016/S1389-1286(00)00179-1.
- [65] Andrew S. Grimshaw. “An Introduction to Parallel Object-Oriented Programming with Mentat”. In: (Apr. 1991).
- [66] Tobias Grosser, Hongbin Zheng, Raghesh Aloor, Andreas Simbürger, Armin Größlinger, and Louis-Noël Pouchet. “Polly - Polyhedral optimization in LLVM”. In: *Proceedings of the First International Workshop on Polyhedral Compilation Techniques (IMPACT '11)* (2011), None.

- [67] Pradeep Kumar Gunda, Lenin Ravindranath, Chandramohan a Thekkath, Yuan Yu, and Li Zhuang. “Nectar : Automatic Management of Data and Computation in Datacenters”. In: *Technology* (2010), pp. 1–8.
- [68] NJ Gunther. “A New Interpretation of Amdahl’s Law and Geometric Scalability”. In: *arXiv preprint cs/0210017* (2002).
- [69] NJ Gunther. “A simple capacity model of massively parallel transaction systems”. In: *CMG-CONFERENCE-* (1993).
- [70] NJ Gunther. “Understanding the MP effect: Multiprocessing in pictures”. In: *In other words* (1996).
- [71] JL Gustafson. “Reevaluating Amdahl’s law”. In: *Communications of the ACM* (1988).
- [72] B Hackett and S Guo. “Fast and precise hybrid type inference for JavaScript”. In: *ACM SIGPLAN Notices* (2012).
- [73] Philipp Haller and Martin Odersky. “Actors That Unify Threads and Events”. In: *Coordination 2007, Lncs 4467* (2007), pp. 171–190. DOI: 10.1007/978-3-540-72794-1{_}10.
- [74] Biao Han, Zhongzhi Luan, Danfeng Zhu, Yinan Ren, Ting Chen, Yongjian Wang, and Zhongxin Wu. “An improved staged event driven architecture for master-worker network computing”. English. In: *CyberC 2009 - International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery i* (Oct. 2009), pp. 184–190. DOI: 10.1109/CYBERC.2009.5342202.
- [75] P.B. Hansen and J. Staunstrup. “Specification and Implementation of Mutual Exclusion”. English. In: *IEEE Transactions on Software Engineering* SE-4.5 (Sept. 1978), pp. 365–370. DOI: 10.1109/TSE.1978.233856.
- [76] Tim Harris, James Larus, and Ravi Rajwar. “Transactional Memory, 2nd edition”. en. In: *Synthesis Lectures on Computer Architecture* 5.1 (Dec. 2010), pp. 1–263. DOI: 10.2200/S00272ED1V01Y201006CAC011.
- [77] Williams Ludwell Harrison. “The interprocedural analysis and automatic parallelization of Scheme programs”. In: *Lisp and Symbolic Computation* 2.3-4 (Oct. 1989), pp. 179–396. DOI: 10.1007/BF01808954.
- [78] CT Haynes, DP Friedman, and M Wand. “Continuations and coroutines”. In: *... of the 1984 ACM Symposium on ...* (1984).
- [79] B He, M Yang, Z Guo, R Chen, and B Su. “Comet: batched stream processing for data intensive distributed computing”. In: *... on Cloud computing* (2010).

- [80] Danny Hendler, Nir Shavit, and Lena Yerushalmi. “A scalable lock-free stack algorithm”. In: *Proceedings of the sixteenth annual ACM symposium on Parallelism in algorithms and architectures - SPAA '04*. New York, New York, USA: ACM Press, June 2004, p. 206. DOI: 10.1145/1007912.1007944.
- [81] M. Herlihy. “A methodology for implementing highly concurrent data structures”. In: *ACM SIGPLAN Notices* 25.3 (Mar. 1990), pp. 197–206. DOI: 10.1145/99164.99185.
- [82] Maurice Herlihy. “Wait-free synchronization”. In: *ACM Transactions on Programming Languages and Systems* 13.1 (Jan. 1991), pp. 124–149. DOI: 10.1145/114005.102808.
- [83] Maurice P. Herlihy. “Impossibility and universality results for wait-free synchronization”. In: *Proceedings of the seventh annual ACM Symposium on Principles of distributed computing - PODC '88*. New York, New York, USA: ACM Press, Jan. 1988, pp. 276–290. DOI: 10.1145/62546.62593.
- [84] C Hewitt, P Bishop, and R Steiger. “A universal modular actor formalism for artificial intelligence”. In: *Proceedings of the 3rd international joint conference on Artificial intelligence* (1973).
- [85] Carl Hewitt. “Viewing control structures as patterns of passing messages”. In: *Artificial intelligence* (1977).
- [86] Carl Hewitt and Jr Baker Henry. “Actors and Continuous Functionals,” in: (Dec. 1977).
- [87] Martin Hilbert and Priscila López. “The world’s technological capacity to store, communicate, and compute information.” In: *Science (New York, N.Y.)* 332.6025 (Apr. 2011), pp. 60–65. DOI: 10.1126/science.1200970.
- [88] C. A. R. Hoare. “Communicating sequential processes”. In: *Communications of the ACM* 21.8 (Aug. 1978), pp. 666–677. DOI: 10.1145/359576.359585.
- [89] C. A. R. Hoare. “Monitors: an operating system structuring concept”. In: *Communications of the ACM* 17.10 (Oct. 1974), pp. 549–557. DOI: 10.1145/355620.361161.
- [90] R D Hornung and J A Keasler. “The RAJA Portability Layer : Overview and Status”. In: (2014).
- [91] YW Huang, F Yu, C Hang, and CH Tsai. “Securing web application code by static analysis and runtime protection”. In: *Proceedings of the 13th ...* (2004).

- [92] Paul Hudak, Thomas Johnsson, Dick Kieburtz, Rishiyur Nikhil, Will Partain, John Peterson, Simon Peyton Jones, Philip Wadler, Brian Boutel, Jon Fairbairn, Joseph Fasel, Maria M. Guzman, Kevin Hammond, and John Hughes. “Report on the programming language Haskell”. In: *ACM SIGPLAN Notices* 27.5 (May 1992), pp. 1–164. DOI: 10.1145/130697.130699.
- [93] John Hughes. “Why functional programming matters”. In: *The computer journal* 32.April 1989 (1989), pp. 1–23. DOI: 10.1093/comjnl/32.2.98.
- [94] Walter Hürsch and Cristina Videira Lopes. *Separation of Concerns*. Tech. rep. NU-CCS-95-03. 1995.
- [95] M Isard, M Budiu, Y Yu, A Birrell, and D Fetterly. “Dryad: distributed data-parallel programs from sequential building blocks”. In: *ACM SIGOPS Operating ...* (2007).
- [96] D Jang and KM Choe. “Points-to analysis for JavaScript”. In: *Proceedings of the 2009 ACM symposium on Applied ...* (2009).
- [97] Kirk L. Johnson, M. Frans Kaashoek, and Deborah A. Wallach. “CRL: High-Performance All-Software Distributed Shared Memory”. In: *ACM SIGOPS Operating Systems Review* 29.5 (Dec. 1995), pp. 213–226. DOI: 10.1145/224057.224073.
- [98] Ralph E. Johnson and Brian Foote. “Designing Reusable Classes Abstract Designing Reusable Classes”. In: *Journal of Object-Oriented Programming* 1 (1988), pp. 22–35.
- [99] N Jovanovic, C Kruegel, and E Kirda. “Pixy: A static analysis tool for detecting web application vulnerabilities”. In: *Security and Privacy, 2006 ...* (2006).
- [100] Gilles Kahn. “The semantics of a simple language for parallel programming”. In: *In Information Processing’74: Proceedings of the IFIP Congress 74* (1974), pp. 471–475.
- [101] Gilles Kahn and David Macqueen. *Coroutines and Networks of Parallel Processes*. en. Tech. rep. 1976, p. 20.
- [102] Hartmut Kaiser. “HPX – A Task Based Programming Model in a Global Address Space”. In: *PGAS 2014*. New York, New York, USA: ACM Press, Oct. 2014, pp. 1–11. DOI: 10.1145/2676870.2676883.
- [103] Hartmut Kaiser, Thomas Heller, and Daniel Bourgeois. *Higher-level Parallelization for Local and Distributed Asynchronous Task-Based Programming*. 2015.

- [104] MN Krohn, E Kohler, and MF Kaashoek. “Events Can Make Sense.” In: *USENIX Annual Technical Conference* (2007).
- [105] L Lamport. “Time, clocks, and the ordering of events in a distributed system”. In: *Communications of the ACM* (1978).
- [106] Leslie Lamport. “Concurrent reading and writing”. In: *Communications of the ACM* 20.11 (Nov. 1977), pp. 806–811. DOI: 10.1145/359863.359878.
- [107] Leslie Lamport, Robert Shostak, and Marshall Pease. “The Byzantine Generals Problem”. In: *ACM Transactions on Programming Languages and Systems* 4.3 (July 1982), pp. 382–401. DOI: 10.1145/357172.357176.
- [108] Charles E. Leiserson. “The Cilk++ concurrency platform”. In: *Journal of Supercomputing* 51.3 (Mar. 2010), pp. 244–257. DOI: 10.1007/s11227-010-0405-3.
- [109] Feng Li, Antoniu Pop, and Albert Cohen. “Automatic Extraction of Coarse-Grained Data-Flow Threads from Imperative Programs”. English. In: *IEEE Micro* 32.4 (July 2012), pp. 19–31. DOI: 10.1109/MM.2012.49.
- [110] Peng Li and Steve Zdancewic. “Combining events and threads for scalable network services implementation and evaluation of monadic, application-level concurrency primitives”. In: *ACM SIGPLAN Notices* 42.6 (June 2007), p. 189. DOI: 10.1145/1273442.1250756.
- [111] B Liskov and L Shrira. *Promises: linguistic support for efficient asynchronous procedure calls in distributed systems*. 1988.
- [112] Dionysios Logothetis, Christopher Olston, Benjamin Reed, Kevin C. Webb, and Ken Yocum. “Stateful bulk processing for incremental analytics”. In: *International Conference on Management of Data* (2010), pp. 51–62. DOI: 10.1145/1807128.1807138.
- [113] S Maffeis, JC Mitchell, and A Taly. “An operational semantics for JavaScript”. In: *Programming languages and systems* (2008).
- [114] S Maffeis, JC Mitchell, and A Taly. “Isolating JavaScript with filters, rewriting, and wrappers”. In: *Computer Security—ESORICS 2009* (2009).
- [115] Nicholas D Matsakis. “Parallel Closures A new twist on an old idea”. In: *HotPar’12 Proceedings of the 4th USENIX conference on Hot Topics in Parallelism* (2012), pp. 5–5.
- [116] Christophe Mauras. “Alpha : un langage equationnel pour la conception et la programmation d’architectures paralleles synchrones”. PhD thesis. Jan. 1989.

- [117] MD McCool. “Structured parallel programming with deterministic patterns”. In: *Proceedings of the 2nd USENIX conference on Hot ...* (2010).
- [118] M Migliavacca and D Eyers. “SEEP: scalable and elastic event processing”. In: *Middleware’10 Posters ...* (2010).
- [119] R. Milner, Mads Tofte, Robert Harper, and David MacQueen. *The Definition of Standard ML - Revised*. 1997, p. 128.
- [120] Derek G. Murray, Frank McSherry, Rebecca Isaacs, Michael Isard, Paul Barham, and Martin Abadi. “Naiad”. In: *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles - SOSP ’13* (Nov. 2013), pp. 439–455. DOI: 10.1145/2517349.2522738.
- [121] Dmitry Namiot and Manfred Sneps-Sneppé. *On Micro-services Architecture*. en. Aug. 2014.
- [122] Jay Nelson. “Structured programming using processes”. In: *Proceedings of the 2004 ACM SIGPLAN workshop on Erlang - ERLANG ’04*. New York, New York, USA: ACM Press, Sept. 2004, pp. 54–64. DOI: 10.1145/1022471.1022480.
- [123] R Nelson. “Including queueing effects in Amdahl’s law”. In: *Communications of the ACM* (1996).
- [124] Leonardo Neumeyer, Bruce Robbins, Anish Nair, and Anand Kesari. “S4: Distributed stream computing platform”. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2010, pp. 170–177. DOI: 10.1109/ICDMW.2010.172.
- [125] Jens Nicolay. “Automatic Parallelization of Scheme Programs using Static Analysis”. PhD thesis. 2010.
- [126] Robert W. Numrich and John Reid. “Co-array Fortran for parallel programming”. In: *ACM SIGPLAN Fortran Forum* 17.2 (Aug. 1998), pp. 1–31. DOI: 10.1145/289918.289920.
- [127] Martin Odersky, Philippe Altherr, Vincent Cremet, Burak Emir, Sebastian Maneth, Stéphane Micheloud, Nikolay Mihaylov, Michel Schinz, Erik Stenman, and Matthias Zenger. “An Overview of the Scala Programming Language”. In: *System Section 2* (2004), pp. 1–130.

- [128] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. “Pig Latin: A Not-So-Foreign Language for Data Processing”. In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08* (June 2008), p. 1099. DOI: 10.1145/1376616.1376726.
- [129] Vivek S Pai, Peter Druschel, and Willy Zwaenepoel. *Flash : An Efficient and Portable Web Server*. 1999. DOI: 10.1.1.119.6738.
- [130] D. L. Parnas. “On the criteria to be used in decomposing systems into modules”. In: *Communications of the ACM* 15.12 (1972), pp. 1053–1058. DOI: 10.1145/361598.361623.
- [131] R Power and J Li. “Piccolo: Building Fast, Distributed Programs with Partitioned Tables.” In: *OSDI* (2010).
- [132] Z Qian, Y He, C Su, Z Wu, and H Zhu. “Timestream: Reliable stream computation in the cloud”. In: *Proceedings of the 8th ACM European Conference on Computer Systems (EuroSys '13)* (2013).
- [133] C Radoi, SJ Fink, R Rabbah, and M Sridharan. “Translating imperative code to MapReduce”. In: *Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages and Applications* (2014).
- [134] Jonathan Ragan-Kelley, Andrew Adams, Sylvain Paris, Frédo Durand, Connelly Barnes, and Saman Amarasinghe. “Halide: A Language and Compiler for Optimizing Parallelism, Locality, and Recomputation in Image Processing Pipelines”. In: *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation* (2013), pp. 519–530. DOI: 10.1145/2491956.2462176.
- [135] Arunmoezhi Ramachandran and Neeraj Mittal. “A Fast Lock-Free Internal Binary Search Tree”. In: *Proceedings of the 2015 International Conference on Distributed Computing and Networking - ICDCN '15*. New York, New York, USA: ACM Press, Jan. 2015, pp. 1–10. DOI: 10.1145/2684464.2684472.
- [136] K.H. Randall. “Cilk: Efficient Multithreaded Computing”. PhD thesis. 1998.
- [137] Veselin Raychev, Martin Vechev, and Manu Sridharan. “Effective Race Detection for Event-driven Programs”. In: *SIGPLAN Not.* 48.10 (Nov. 2013), pp. 151–166. DOI: 10.1145/2544173.2509538.
- [138] DP Reed. “” Simultaneous” Considered Harmful: Modular Parallelism.” In: *HotPar* (2012).

- [139] J Rees and W Clinger. “Revised report on the algorithmic language scheme”. In: *ACM SIGPLAN Notices* 21.12 (Dec. 1986), pp. 37–79. DOI: 10.1145/15042.15043.
- [140] MC Rinard and PC Diniz. “Commutativity analysis: A new analysis framework for parallelizing compilers”. In: *ACM SIGPLAN Notices* (1996).
- [141] H. Sackman, W. J. Erikson, and E. E. Grant. “Exploratory Experimental Studies Comparing Online and Offline Programming Performance”. In: *Commun* 11.1 (1968), pp. 3–11. DOI: 10.1145/362851.362858.
- [142] Tiago Salmito, Ana Lucia de Moura, and Noemi Rodriguez. “A Flexible Approach to Staged Events”. English. In: *2013 42nd International Conference on Parallel Processing* (Oct. 2013), pp. 661–670. DOI: 10.1109/ICPP.2013.80.
- [143] Tiago Salmito, Ana Lúcia de Moura, and Noemi Rodriguez. “A stepwise approach to developing staged applications”. In: *The Journal of Supercomputing* (Jan. 2014). DOI: 10.1007/s11227-014-1110-4.
- [144] O. Shivers. “Control-flow analysis of higher-order languages”. PhD thesis. 1991, pp. 1–186.
- [145] Herbert A. Simon. “The architecture of complexity”. In: *Proceedings of the American Philosophical Society* 6.106 (1962), pp. 467–482. DOI: 10.1109/MCS.2007.384127. arXiv: 0205649 [cond-mat].
- [146] Elliott Slaughter, Wonchan Lee, Sean Treichler, and Michael Bauer. “Regent : A High-Productivity Programming Language for HPC with Logical Regions”. In: *SC* (2015). DOI: 10.1145/2807591.2807629.
- [147] GD Smith. “Local reasoning about web programs”. In: (2011).
- [148] M Sridharan, J Dolby, and S Chandra. “Correlation tracking for points-to analysis of JavaScript”. In: *ECOOP 2012–Object- ...* (2012).
- [149] W. P. Stevens, G. J. Myers, and L. L. Constantine. “Structured design”. English. In: *IBM Systems Journal* 13.2 (1974), pp. 115–139. DOI: 10.1147/sj.132.0115.
- [150] B Stroustrup. “The C++ programming language”. In: (1986).
- [151] Kevin J. Sullivan, William G. Griswold, Yuanfang Cai, and Ben Hallen. “The structure and value of modularity in software design”. In: *ACM SIGSOFT Software Engineering Notes* 26.5 (Sept. 2001), p. 99. DOI: 10.1145/503271.503224.

- [152] H. Sundell and P. Tsigas. “Fast and lock-free concurrent priority queues for multi-thread systems”. In: *Proceedings International Parallel and Distributed Processing Symposium* 00.C (2003), p. 11. DOI: 10.1109/IPDPS.2003.1213189.
- [153] Gerald Jay Sussman and Jr Steele, Guy L. “Scheme: A interpreter for extended lambda calculus”. In: *Higher-Order and Symbolic Computation* 11 (1998), pp. 405–439. DOI: 10.1023/A:1010035624696.
- [154] Richard E Sweet. “The Mesa programming environment”. In: *ACM SIGPLAN Notices*. Vol. 20. 7. 1985, pp. 216–229. DOI: 10.1145/17919.806843.
- [155] David Tarditi, Sidd Puri, and Jose Oglesby. “Accelerator: using data parallelism to program GPUs for general-purpose uses”. In: *Proceedings of the 12th international conference on Architectural support for programming languages and operating systems* 34.5 (Oct. 2006), pp. 325–335. DOI: 10.1145/1168918.1168898.
- [156] P. Tarr, H. Ossher, W. Harrison, and Jr. Sutton, S.M. “N degrees of separation: multi-dimensional separation of concerns”. In: *Proceedings of the 1999 International Conference on Software Engineering (IEEE Cat. No.99CB37002)* (1999), pp. 107–119. DOI: 10.1145/302405.302457.
- [157] William Thies, Michal Karczmarek, and Saman Amarasinghe. “StreamIt: A language for streaming applications”. In: *Compiler Construction* LNCS 2304 (2002), pp. 179–196. DOI: 10.1007/3-540-45937-5.
- [158] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. “Hive”. In: *Proceedings of the VLDB Endowment* 2.2 (Aug. 2009), pp. 1626–1629. DOI: 10.14778/1687553.1687609.
- [159] Shahar Timnat, Anastasia Braginsky, Alex Kogan, and Erez Petrank. “Wait-free linked-lists”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7702 LNCS. 2012, pp. 330–344. DOI: 10.1007/978-3-642-35476-2_{_}23.
- [160] Ankit Toshniwal, Jake Donham, Nikunj Bhagat, Sailesh Mittal, Dmitriy Ryaboy, Siddarth Taneja, Amit Shukla, Karthik Ramasamy, Jignesh M. Patel, Sanjeev Kulkarni, Jason Jackson, Krishna Gade, and Maosong Fu. “Storm@ twitter”. In: *Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD ’14*. New York, New York, USA: ACM Press, June 2014, pp. 147–156. DOI: 10.1145/2588555.2595641.

- [161] Konrad Trifunovic, Albert Cohen, David Edelsohn, Feng Li, Tobias Grosser, Harsha Jagasia, Razya Ladelsky, Sebastian Pop, Jan Sjödin, and Ramakrishna Upadrasta. *GRAPHITE Two Years After: First Lessons Learned From Real-World Polyhedral Compilation*. en. Jan. 2010.
- [162] D Turner. “An overview of Miranda”. In: *ACM SIGPLAN Notices* 21.12 (Dec. 1986), pp. 158–166. DOI: 10.1145/15042.15053.
- [163] D. A. Turner. “The semantic elegance of applicative languages”. In: *Proceedings of the 1981 conference on Functional programming languages and computer architecture - FPCA '81*. New York, New York, USA: ACM Press, Oct. 1981, pp. 85–92. DOI: 10.1145/800223.806766.
- [164] Gautam Upadhyaya, Vijay S. Pai, and Samuel P. Midkiff. “Expressing and exploiting concurrency in networked applications with aspen”. In: *Proceedings of the 12th ACM SIGPLAN symposium on Principles and practice of parallel programming - PPOPP '07* (Mar. 2007), p. 13. DOI: 10.1145/1229428.1229433.
- [165] John D. Valois. “Lock-free linked lists using compare-and-swap”. In: *Proceedings of the fourteenth annual ACM symposium on Principles of distributed computing - PODC '95*. New York, New York, USA: ACM Press, Aug. 1995, pp. 214–222. DOI: 10.1145/224964.224988.
- [166] Hans Vandierendonck, Sean Rul, and Koen De Bosschere. “The Parallax infrastructure: automatic parallelization with a helping hand”. In: *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*. New York, New York, USA: ACM Press, Sept. 2010, pp. 389–399. DOI: 10.1145/1854273.1854322.
- [167] Philip Wadler. “The essence of functional programming”. In: *Proceedings of the 19th ACM SIGPLAN-SIGACT symposium on Principles of programming languages - POPL '92*. New York, New York, USA: ACM Press, Feb. 1992, pp. 1–14. DOI: 10.1145/143165.143169.
- [168] M Wand. “Continuation-based multiprocessing”. In: *Proceedings of the 1980 ACM conference on LISP and ...* (1980).
- [169] S Wei and BG Ryder. “State-sensitive points-to analysis for the dynamic behavior of JavaScript objects”. In: *ECOOP 2014–Object-Oriented Programming* (2014).
- [170] M Welsh, D Culler, and E Brewer. “SEDA: an architecture for well-conditioned, scalable internet services”. In: *ACM SIGOPS Operating Systems Review* (2001).

- [171] Martin Wimmer, Jakob Gruber, Jesper Larsson Träff, and Philippas Tsigas. “The lock-free k-LSM relaxed priority queue”. In: *Proceedings of the 20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming - PPOPP 2015*. New York, New York, USA: ACM Press, Jan. 2015, pp. 277–278. DOI: 10.1145/2688500.2688547.
- [172] Sunny Wong, Yuanfang Cai, Giuseppe Valetto, Georgi Simeonov, and Kanwarpreet Sethi. “Design Rule Hierarchies and Parallelism in Software Development Tasks”. In: *2009 IEEE/ACM International Conference on Automated Software Engineering*. IEEE, Nov. 2009, pp. 197–208. DOI: 10.1109/ASE.2009.53.
- [173] Reynold S. Xin, Josh Rosen, Matei Zaharia, Michael J. Franklin, Scott Shenker, and Ion Stoica. “Shark”. In: *Proceedings of the 2013 international conference on Management of data - SIGMOD ’13*. New York, New York, USA: ACM Press, June 2013, p. 13. DOI: 10.1145/2463676.2465288.
- [174] Sunghwan Yoo, Hyojeong Lee, Charles Killian, and Milind Kulkarni. “InContext : Simple Parallelism for Distributed Applications Categories and Subject Descriptors”. In: *HPDC*. New York, New York, USA: ACM Press, June 2011, p. 97. DOI: 10.1145/1996130.1996144.
- [175] D Yu, A Chander, N Islam, and I Serikov. “JavaScript instrumentation for browser security”. In: *ACM SIGPLAN Notices* (2007).
- [176] Yuan Yu, Michael Isard, Dennis Fetterly, Mihai Budiu, Ulfar Erlingsson, Pradeep Kumar Gunda, Jon Currey, Frank McSherry, Kannan Achan, and Christophe Poulain. “Some sample programs written in DryadLINQ”. In: *Microsoft Research* (2009).
- [177] Tomofumi Yuki, Gautam Gupta, Daegon Kim, Tanveer Pathan, and Sanjay Rajopadhye. “AlphaZ: A system for design space exploration in the polyhedral model”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7760 LNCS (2013), pp. 17–31. DOI: 10.1007/978-3-642-37658-0_{_}2.
- [178] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, and Ankur Dave. “Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing”. In: *NSDI’12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation* (Apr. 2012), pp. 2–2. DOI: 10.1111/j.1095-8649.2005.00662.x. arXiv: EECS-2011-82.

- [179] Yili Zheng, Amir Kamil, Michael B. Driscoll, Hongzhang Shan, and Katherine Yelick. “UPC++: A PGAS Extension for C++”. In: *2014 IEEE 28th International Parallel and Distributed Processing Symposium*. 2014, pp. 1105–1114. DOI: 10.1109/IPDPS.2014.115.