JESPER BOEG

PRIMING KANBAN

A 10 step guide to optimizing flow in your software delivery system

FOREWORD BY JAMES SUTTON

InfoQ
ENTERPRISE SOFTWARE DEVELOPMENT SERIES

Trifork Agile Excellence Mini Book Series

software pilots TRIFORK.

Biography: Jesper Boeg

Jesper has worked as an Agile and Lean coach since early 2006, and he is now VP of the department for ”Agile Excellence” at Trifork. He has a Master’s degree from Aalborg University in the area of Information Systems and wrote his thesis on how to successfully manage distributed software teams.

Jesper helps teams and organizations adopt Agile and Lean principles, with a strong focus on understanding “why”. He has a reputation for being honest and straightforward, with a firm belief that change management is much more about people than process.

Jesper believes that trust is best established through an unrelenting focus on transparency in the entire organization. He has a strong passion for Lean Product Development and continuously emphasises that one must look at the entire software delivery system in order to guide success.

Context Based Strategically Aligned Agility are keywords in Jesper’s work. It is his experience that in order to create lasting change, organizations cannot rely on Best Practice rule sets, but rather must put effort into understanding “why” as well as aligning Agile principles with the overall business strategy. Otherwise, they will quickly revert to former practices when faced with difficulty and will restrict themselves from great improvement opportunities.

Jesper regularly speaks at Agile and Lean conferences. He is a member of the GOTO Aarhus Program Advisory Board and has served as a track-host on numerous GOTO and QCon conferences.

JESPER BOEG

PRIMING KANBAN

A 10 step guide to optimizing flow in your software delivery system

Written by Jesper Boeg
Foreword by James Sutton
Designed by Cecilie Marie Skov
Graphics by Cecilie Marie Skov and Jesper Boeg

First edition October 2011, Second edition February 2012
2. printing
Printed in Denmark at Chronografisk A/S

Trifork A/S
Aarhus: Margrethepladsen 4, DK-8000 Aarhus C
Copenhagen: Spotorno Alle 4, DK-2630 Taastrup
Phone: +45 8732 8787
E-mail: info@trifork.com

Thanks to everybody that helped review this mini-book. The readability and the content have been greatly improved as a result of your comments. I would like to give a special thanks to Yuval Yeret, Karl Scotland and James Sutton for your insightful comments as well as taking the time to review the book in such detail.

Contents

Foreword by James Sutton

Kanban is an industrial technique for "pulling" work through its entire lifecycle, causing the work to flow more smoothly and at a higher rate. Kanban also shines a light on the activities that are normally hidden. This visibility is at two levels; for individual activities, and also over the lifecycle as a whole.

For many decades factory production has enjoyed the benefits of kanban: higher productivity, better quality, and more-satisfied workers who enjoyed increased insights into and control over their work. Until recently, however, nobody had figured out how the ideas of kanban might translate into a knowledge work field like software development. It wasn't even clear if it would help in such fields.

In the 2000s, David Anderson worked out the Rosetta Stone for such a translation, and developed the ideas of Kanban into an approach well-suited for software development. His 2010 book "Kanban" painted the picture at, say, VGA-resolution. It set an entire industry experimenting with the Kanban approach in many different product types and situations. They found that Kanban works with any development lifecycle, from Scrum to Waterfall. It complements rather than competes with them.

Anderson's book left two things still wanting, however: An additional level of detail in the theory and practice of knowledge-work Kanban, and a "starter's guide" for people about to take plunge.

The additional needed level of detail is in an upcoming book by David Anderson. You could say that it will be at HD-resolution for display on a big screen TV (I've talked to David some about the book).

The other piece of the puzzle, the "starter's guide," is what you are about to read: Jesper Boeg's mini-book "Priming Kanban." Call it an iPhone video-podcast with which to follow along wherever you do your own work. A great complement to the big HD picture.

When I recently received a draft of Jesper's book for review, I didn't think it possible for such a small book to contain both a concise overview and a practical, step-by-step worker's introduction. As I read along, though, I realized that Jesper had succeeded at both.
He has a unique gift for getting to the core of a matter, and then helping others get to the core of it for themselves.

I heartily recommend "Priming Kanban" as a practitioner's quick-start introduction to using Kanban. It will enable you to try out Kanban more quickly and painlessly…and defuse your anxiety about doing so. On a broader scale, as it inspires many people to take the same plunge, our industry will expand the use of this newest tool in our toolkit, Kanban…with all the benefits it brings.

And that's good for all of us.

James Sutton

CEO, The Jubata Group
President, Lean Software & Systems Consortium
Shingo Prize, 2007
INCOSE ESEP

Introduction

Background

Before we dive into the step-by-step guide to implementing Kanban, let us spend a few minutes introducing the concept so that you will be able to recognise where each step fits into the overall Kanban change management framework. The scope of this mini book is not to describe Kanban concepts in depth; for that, I refer to David J. Anderson's excellent book Kanban, which I strongly recommend reading:

http://www.amazon.com/Kanban-Successful-Evolutionary-Technology-Business/dp/0984521402/ref=sr_1_1?ie=UTF8&qid=1313588404&sr=8-1

Instead, I hope to give a short introduction, followed by step-by-step advice on how to get started.

Kanban, or more precisely "Kanban system for software development", represents a more direct implementation of Lean Product Development principles in software development, compared to traditional Agile methods. With a consistent focus on flow and context, Kanban offers a less prescriptive approach to Agile and has become a popular extension to traditional methods, like Scrum and XP.

The word "Kanban" is Japanese and means "Visual Card". The reason that Google returns more than 5 million results on a search for Kanban is, however, that it also used to describe the system that has been used at Toyota for decades to visually control and balance the production line and which has become almost synonymous with the implementation of Lean principles. Therefore, while kanban systems are a relatively new concept in IT, it has been used for more than 50 years in Lean production systems at Toyota.

The use of Kanban in software was pioneered by David Anderson, who, in close collaboration with Don Reinertsen, has strived to expand the knowledge of Lean and the use of Kanban to visualize and optimize the workflow in IT development, maintenance and operations.

When should I consider working with Kanban?

If the answer is yes to one or more of the following questions, there is a good chance you will benefit from reading the rest of this book:

- Have you been struggling with implementing Agile in your organiza tion for a while without much success?
- Have you been using Agile for a while and performance improve ments have started to level off?
- Are you using valuable time on Agile practices, which no longer seems to fit the context you are working in?
- Have you been using Agile as a checklist without fully understading the underlying principles?
- Do you have a need for more flexibility than frozen, committed and planned iterations have to offer?
- Do your priorities shift on a daily basis?
- Are you using processes designed for Agile product development in a context where they are not an easy fit, e.g. maintenance and operations?
- Do you need a gradual transition from waterfall type execution to Agile in order to avoid high levels of organizational resistance?

Despite whether your goal is to work in a strict Scrum context, if you are using waterfall or are trying to find a way to super optimize your current Agile implementation, most will benefit from the deeper understanding of Lean, which Kanban has proven to be an excellent catalyst for.

What is Kanban?

There are a variety of approaches to Kanban, but most agree that Kanban is a change management method that focuses on the following principles:

- **Visualize Work**
  - Visualize every step in your value chain from vague concept to releasable software.
- **Limit Work-In-Progress (WIP)**
  - Set explicit limits on the amount of work allowed in each stage.
- **Make Policies Explicit**
  - Make the policies you are acting according to explicit.
- **Measure and Manage Flow**
  - Measure and Manage Flow to make informed decisions and visualise consequence
- **Identify Improvement Opportunities**
  - Create a Kaizen culture where continuous improvement is everyone's job.

With the underlying philosophy that you should:

- **Start with what you do now**
- **Agree to pursue incremental, evolutionary change**
- **Respect the current process, roles, responsibilities & titles**

Those of you familiar with Lean will recognize many of these principles as being the foundation for a Lean pull system and the building of a continuous improvement (Kaizen) culture. What Kanban first and foremost does is serve as a catalyst to introduce Lean ideas into software delivery systems. This is also stated in David's book:

"...Kanban (capital K) is the evolutionary change method that utilizes a kanban (small k) pull system, visualization, and other tools to catalyze the introduction of Lean ideas into technology development and IT operations"

David J. Anderson, Kanban 2010

We will show a lot of examples of Kanban boards in the following chapters and will explain the mechanics. To give you an idea of the concept, figure 1 shows an example.

As you can see, all work is made visible. WIP limits are in place (written in each column header). Policies are made explicit and flow is measured. As you will notice, the last column does not have a WIP limit assigned. This is due to the fact that this particular team has opted for a regular weekly release cadence (3pm Tuesday), which means that all finished work is released at this time.

Kanban is all about driving evolutionary change and these simple steps have proven extremely helpful in doing that. The reason we refer to such a system as a "Kanban Pull System" is that visualized flow and WIP limits ensure that you can never introduce more work into the system than it has capacity to handle. There is only a certain number of work permits (kanbans) available, so you must complete existing work before new work can be started. This results in functionality being pulled through the system, based on capacity rather than pushed based on forecasts or demand.

There are no rules in terms of how your board should look. The only limitation is your imagination, creativity and the constraints of an electronic system or wall space. Since this is a starter's guide, we will, however, use pretty basic board examples to demonstrate the principles.

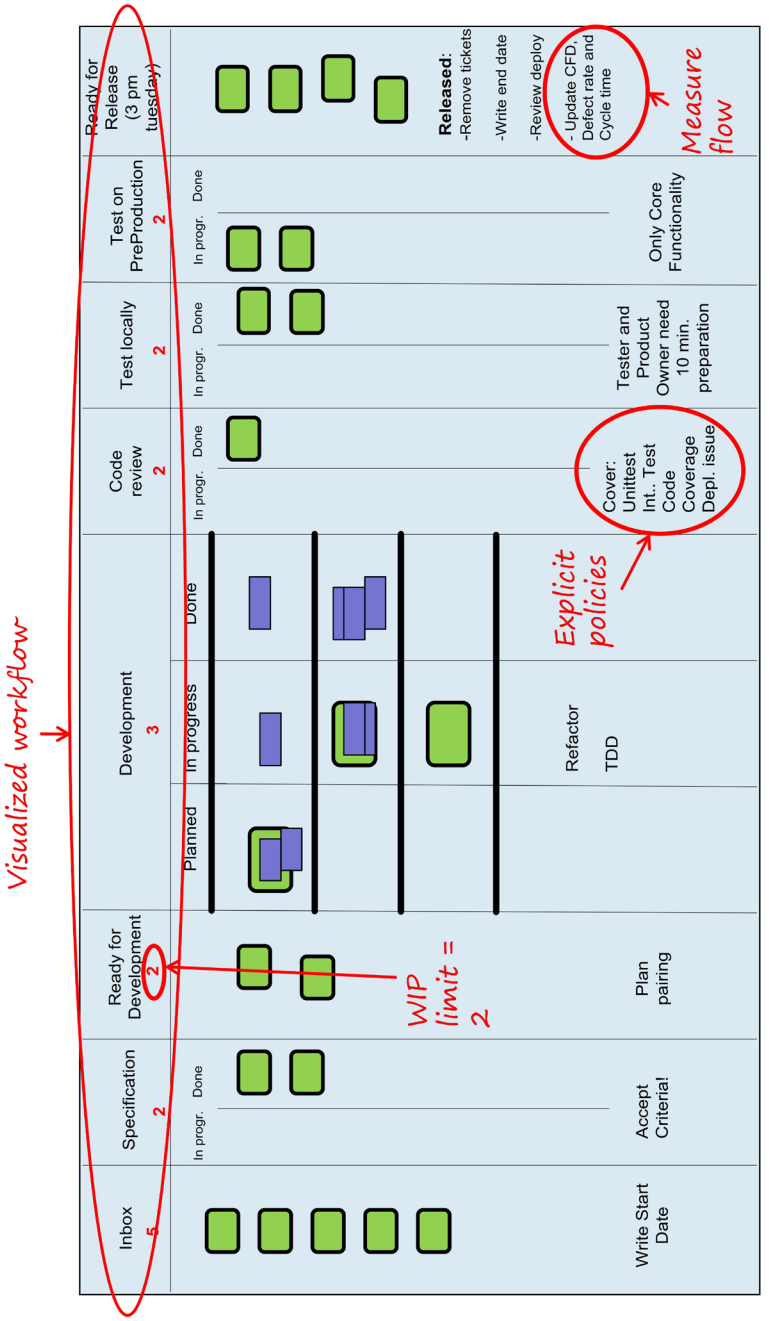

Fig. 1 Kanban Principles in Action

How do we get started with Kanban?

Hopefully, the 10 steps in this book will get you well on your way, but before we get that far, it is important to understand that Kanban has a different approach to change management than most other Agile methods.

Kanban is built on the concept of evolutionary change. Start by understanding how your current software delivery system works. When you have managed to visualize, measure and manage your flow, improve it one step at a time by relieving the largest bottleneck. This is quite different compared to e.g. Scrum, where you will often start out by redefining roles, process and artefacts. This makes Kanban ideally fitted for use on top of existing processes, which can be anything from Scrum to Waterfall and perfect in situations where organizational structures inhibit radical change. Remember the foundational principle:

"Respect the current process, roles, responsibilities & titles"

In Lean terms, this means that Kanban is primarily built on the concept of Kaizen (continuous improvement), and Kaikaku (dramatic change) is only used in special situations where structural change is needed or where serious performance leverage needs to be identified.

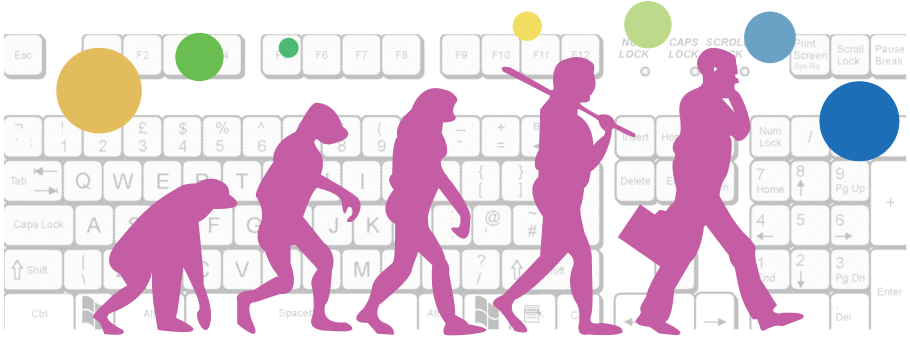

Fig. 2 Kanban Takes an Evolutionary Approach to Process Optimization

Where can Kanban be used?

Now we are almost ready to get started. However, before we dive into the implementation details, let us just quickly bust some of the myths about Kanban to make sure that the following sections are read with the right mindset in place.

At Trifork, we have helped a lot of companies and teams increase their effectiveness by adopting Kanban. At first, it seemed that the primary target groups were teams working with maintenance and operation, but Kanban has proven to be just as helpful for software development. Moreover, teams and organizations working with waterfall-like methods have found the evolutionary approach to be extraordinarily helpful in a gradual transition to Agile product development.

Kanban Myths

- **Myth:** : Kanban is only suitable for teams working with small uniform tasks, like those seen in operation and maintenance.
- **Fact:** Kanban is heavily inspired by Don Reinertsen's work with Lean Product Development and has proven to be as good a fit to software development as it has been for operation and maintenance.

- **Myth:** Kanban and Scrum are opposites.
- **Fact:** None of the principles in Kanban restrict you from doing Scrum. Kanban acts as a change agent and the principles in Scrum should therefore only be used in cases where they help to optimize flow. Nothing is keeping you from starting with Scrum and using Kanban to drive further change – many projects have been incredibly successful with this strategy. Some might even argue that it was the original intention with Scrum as well; however, it somehow got lost in the focus on ceremonies, roles and artifacts.

- **Myth:** By not insisting on planned committed iterations, Kanban falls prey to Parkinson's Law that "Work expands so as to fill the time available for its completion".
- **Fact:** Though being a valid concern, Kanban projects rarely display this behaviour, since fixed cadences, extreme visualization,

cycle time measurement and tighter feedback loops with business keep focus tight and work items flowing.

- ***Myth:*** Kanban teams do not use timeboxes.
- **Fact:** Timeboxes are not mandatory, but should be used when they help optimize flow, feedback and quality. Most Kanban teams use fixed but decoupled cadences of planning, review and releases, and thereby dispense with the traditional iteration model, while keeping the value intact.

- **Myth:** Kanban teams do not estimate.
- **Fact:** Estimates are not mandatory, but should be used when appropriate. Most Kanban development projects use some degree of initial sizing to ensure optimal portfolio management, prioritization and alignment, or use estimates for work that is more sensitive to cost/benefit analysis or due date performance.

- **Myth:**Kanban is better than Scrum/XP/Crystal/FDD….
- **Fact:** Kanban is first and foremost a catalyst for driving change and therefore needs a starting point.

Therefore, while most projects will benefit from using Kanban, it is not a substitute for e.g. Scrum. Scrum is in most cases a perfect starting point when adopting Kanban.

Though the Kanban community is constantly fighting these and other myths, Kanban still remains one of the most misunderstood concepts in the Agile community to this date for two main reasons:

A lot of the noise and confusion revolves around the fact that Kanban is a change method and therefore has very few descriptive parts telling you how to work, which roles to fill, etc. Since the concept of a change method is poorly understood, people have tried to compare it to more prescriptive methods, such as Scrum and XP.

Local examples and emergent behaviours of Kanban, in real world projects, have come to represent a “Kanban method” to some people. It is easy to see why people misunderstand, since many Kanban projects show the same emergent practices. Reality, however, is that Kanban is about using Lean principles to optimize existing processes in an evolutionary way, and therefore cannot and should not be compared to Scrum, XP, Crystal, FDD or whatever method you are using.

The second reason is arguably that the word “Kanban” might carry too much baggage from its origins in Lean production systems to be an adequate word to describe a change method for software development and IT operations. Though kanban pull systems, as they are used in production systems, do drive change, Kanban in software builds on a much broader set of Lean principles, and that creates a difficult mental gap for those having worked extensively with Lean in the past. You will find that a lot of the strong reactions against Kanban, from parts of the Agile community, actually build on this misunderstanding.

The good news is that most projects, Agile or non-Agile, can benefit hugely from using the principles of Kanban to drive change and continuous improvement, which I hope to demonstrate in this book.

The observant reader might have noticed that the title of this book is not exactly in line with the Kanban principles. As David Anderson wrote in a tweet on June 30, 2011 “… Kanban system design is a thinking process, not a copying or template implementing process”. However, those familiar with the “Dreyfus Model of Skill Acquisition” will recognize that whenever acquiring a new skill, you need prescriptions at first; my hope is that this “primer” will help you transition quicker and more painlessly. The important thing is to know that prescriptions only serve as a way for you to gain knowledge to move forward and not as an endpoint or a checklist to be followed blindly. The templates and practices suggested in the 10 steps in this book are emergent behaviours experienced in Kanban projects, and not the Kanban change method itself.

Now that we have got the general picture, let’s go on to how we can apply these things in practice. Each step consists of a short explanation on “why” followed by “how”.

Step 1 Visualize your workflow

The first step towards visualizing your workflow is to understand how your current system works.

Understanding your software delivery system

To be able to make informed decisions about how to best optimize your workflow, the first step is to understand what you are doing. The important thing here is to resist the temptation to change anything. Just find out how you are working without idealizing it. The key is to try to map your entire software delivery workflow and not just focus on the “development” part.

There are a number of different ways to do this. The most popular way is to use the Lean concept of Value Stream Maps (VSM). Recently, VSMs have taken quite a beating from Agile and other knowledge work communities - the main argument being that knowledge work is not a linear process, like the ones seen in production systems. This has led to the evolvement of techniques, like Knowledge Creation Networks, which are better suited to handle non-linear work. For this simple example, we will, however, use the more simple VSM technique, which I still find extraordinarily helpful. However, you should explore the option that fits your context.

In its simplest form, a Value Stream Map is a visualization of the stages that our work passes through, from raw material to finished product, or in the case of software, from vague idea to a feature working in production. The key thing when doing this for knowledge work is to think of each stage as the primary form of information arrival. For example, a stage called “Test” includes more work than just testing (fixing, analyzing, refactoring, discussions, updating, accept criteria, etc.), but since the primary form of information arrival is “Test”, we will define our work as being in the “Test” stage, while all these activities are going on. The space between our stages, where no information is being added, is defined as the “wait time”. Figure 3 shows an example from a real project.

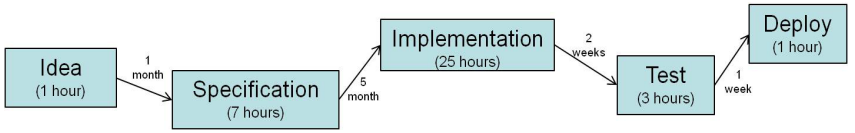

Fig. 3 Value Stream Map Example

We might later decide that implementation actually consists of: User story breakdown, Development and Code Review, but let's stick with this simple version for now. Already, improvement ideas spring into our mind: Why is there an average wait time of 5 months from specification to implementation? Why are we waiting 2 weeks before testing? Resist the temptation to fix these issues right now, there will be plenty of time for that later; right now, our focus is on understanding how our current system works.

Initially, it is a good idea to limit the number of stages in the value stream map and Kanban board. With too many stages in place, you quickly lose sight of the big picture and just focus on the mechanics. Later on, you might find it beneficial to add more details, but keep it simple for now.

Visualizing your system

Now that we have gained a better understanding of our software delivery system, the next step is to try to visualize it. We can do this in an electronic system or by simply using a whiteboard. Unless you are working in a distributed team, it is usually a good idea to start out with just the whiteboard; nothing makes your work more visible than having it right in front of you all the time and being able to physically touch it. As team maturity increases and you find the need to collect more data, you might want to move to an electronic version, but stick to the whiteboard for now.

Often, you will end up with at least two types of stages: "Activity" stages, where active work is being performed, and "Buffer" stages, where work is waiting to be released/developed, etc., but more on that later.

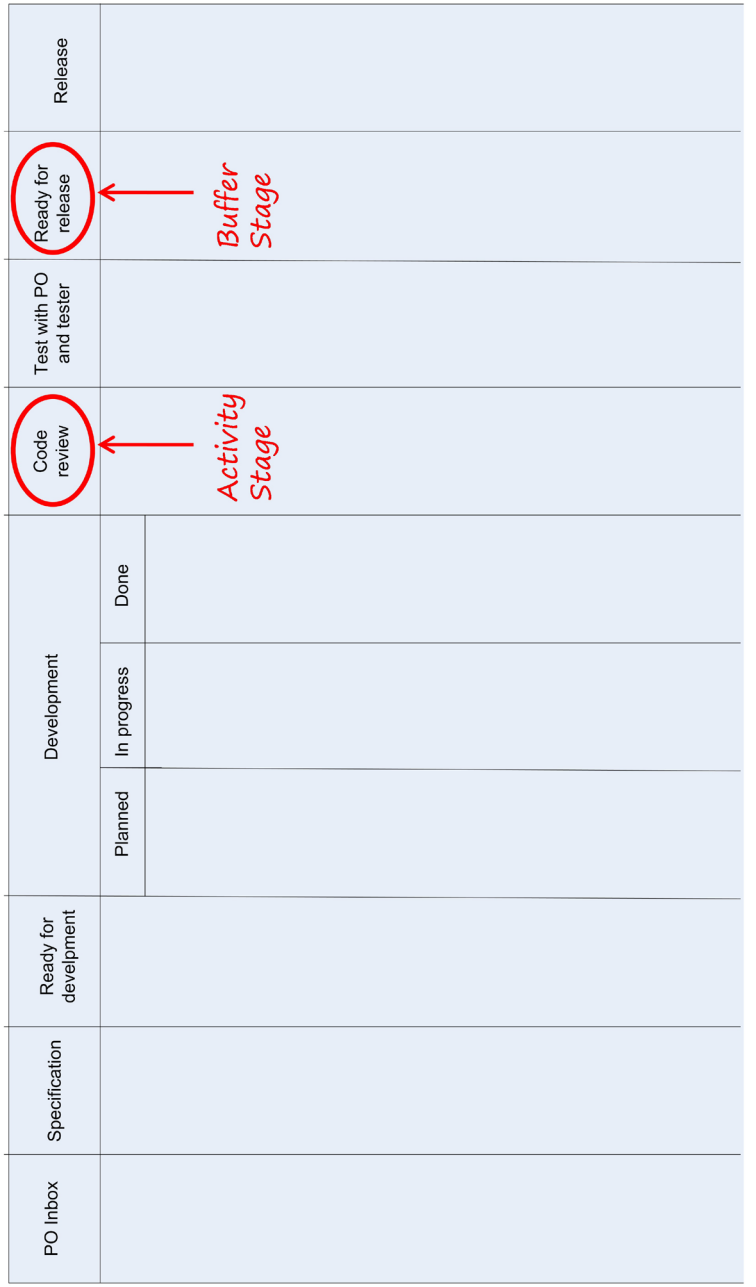

Fig. 4 Activity Stage vs. Buffer Stage

The first version of your board might look something like the one shown in figure 4 (Notice that all the work necessary to complete a given feature is represented, not just development).

Every feature starts out as a vague idea in the PO Inbox and ends up in the "Releasing" column, where it is removed from when the feature is actually working in production. Notice that we haven't changed anything - you might still be following a strict Scrum implementation.

If visualizing your work proves a difficult task, now is the place to stop. Don't go any further before you have managed to visualize all the work you do. If information is hidden and some tasks are completed outside of your workflow system, there is little chance that you will ever be able to make informed decisions about how best to optimize. A general rule is that "you can only manage the work you can see". Visualizing work may sound deceptively simple, but it can prove difficult in real life. Reasons may be varied: People know they are doing things they shouldn't. They are afraid they will be punished if their superiors know how things really work. Though stressed out, they feel they will let their colleagues down if they are not constantly fire fighting. You need to fix these problems and make everybody involved understand that no one will be blamed or discredited for displaying the current status, before moving on.

Visualizing your workflow gives a number of benefits - the most important being:

Focus on "The Whole"

- It becomes visible exactly how your work affects others, and vice versa.

Transparency

- Everybody knows exactly what is going on and no information is hidden.

Identifying waste

- You naturally start to question why you are doing things the way you are (more on that later)

In general, we want work to flow from left to right on our board, since when things start moving backwards, our system becomes much more complicated to master. To do this, we need to accept that each stage on the Kanban board represents the primary activity and that e.g. "test" also includes some amount of analysis, coding, review and documentation. In special cases where "test" reveals an issue that will take 100 hours to fix, you should consider starting from scratch, since the premises have changed dramatically.

Do not restrict your board design to what you have seen others use or to the examples in this book. It seems that every time I help kick-start a team, I see new creative examples and good ideas on how to best visualize flow. The careful reader might even have noticed that the more advanced concept of expanding and collapsing work items sneaked in to the very first board example in this book (figure 1). This is a great example of how teams have found ways to map the behaviour of a user story, expanding into a series of tasks (purple tickets) that move across the three stages of development individually before collapsing again when all tickets have been completed and the user story is reviewed as a whole. It is a simple concept, but is extremely powerful in displaying what is going on.

Step 2 Limit Work in Progress (WIP)

When you have managed to visualize your workflow, you are ready to proceed to the next step - limiting WIP. Though it might be tempting to do this immediately, visualizing work is often not as easy as it seems and it is therefore often a good idea to spend some time exploring this aspect before continuing on.

Understanding WIP

To understand why limiting WIP makes sense, we need to take a look at Little's law, which states that (adapted to product development terminology): Cycle time = WIP / Throughput per unit of time

Cycle Time describes the time it takes for a work item to pass through our system or, in other words, "the time it takes from when a feature is selected for implementation until it is working in production". How you define "selected for implementation" depends on your context. For some, it is the placement of an item on the backlog and for others; it might be the time an item is selected for detailed specification. You might also want to distinguish between the two and refer to the time an item arrives until it is delivered as "Lead Time" and the time from when it is selected for implementation until it is delivered as "Cycle Time".

WIP describes the amount of Work In Progress in our system. How many "story points"/"user stories"/"backlog items" are currently in progress in our system? Again, it depends on the context. Some include all items on the backlog in WIP, while others consider only the items selected for implementation.

Throughput per unit of time is simply the average number of items produced in a given period of time. In Scrum, this is usually referred to as velocity.

This means that given a system with 100 user stories in progress (WIP) and a throughput of 2 user stories per week, the average cycle time is 100/2 = 50 weeks or almost a year. Reducing this to 25 weeks can be done by either doubling throughput to 4 user stories per week or by reducing the number of user stories in progress to 50. In most cases, it is initially much easier to reduce WIP than increasing throughput.

As you might have guessed, limiting WIP is all about reducing the cycle time to increase flow and to minimise the amount of work we have invested time and resources in, but has yet to generate any business value. Fast feedback cycles are also a great way to minimise risk, since decisions are validated continuously and quality issues are exposed immediately. This is a subject explored in detail in Capers Jones' "Cost of Quality" (1980).

So how do we do this? Well actually, all we have to do initially is to make our best effort to define how many items we will allow in each stage of our board at any given time. A good idea is to let this exercise be guided by the policies your team would like to enforce. If the team decides that it is a good idea that no developer should work on a user story single-handedly, you might choose a limit of 3 for a team of 6. Note that this is only true for activity columns, like "development", "test", etc. For buffer columns, like "ready for development", the general rule is that if it is empty once a year, the WIP limit is too large (364 days you are working with a larger buffer than needed) and if it is empty once a day, it is too small. People often joke that the universal WIP limit is 5, so if in doubt, 5 is probably a good number to start with.

Visualizing WIP Limits

How you visualize your limit is up to you. Figures 5 and 6 show two common ways of doing it. In figure 5, only one item may be placed in a box and that gives you a very visual signal of when you have a "permit" to start/pull new work (the box is empty). In figure 6, it is easier to divide the activity stage into "in progress" and "done", since the WIP limit is simply written in each column header. This can give you additional insight into how your system is working and this is the common way of doing it in IT. People having worked with Lean manufacturing might opt for the first version, since it more closely resembles the visual pull signal of the plastic card.

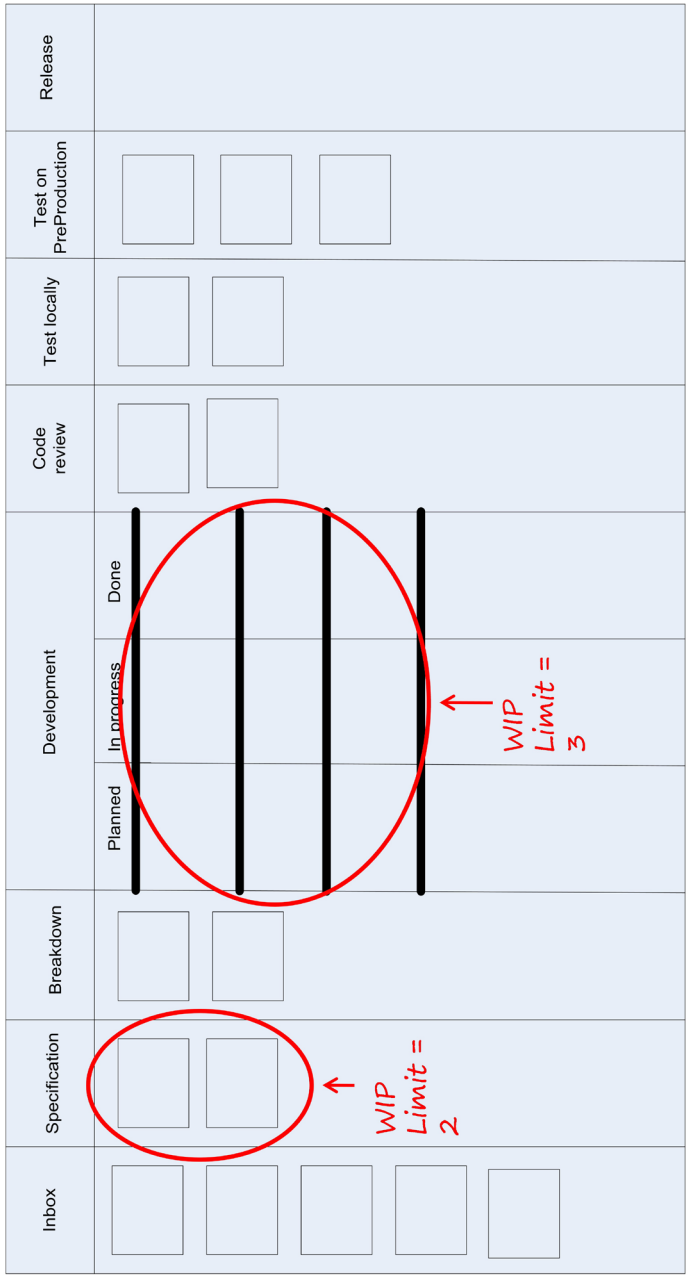

Fig. 5 WIP Limits Visualized Using Containers

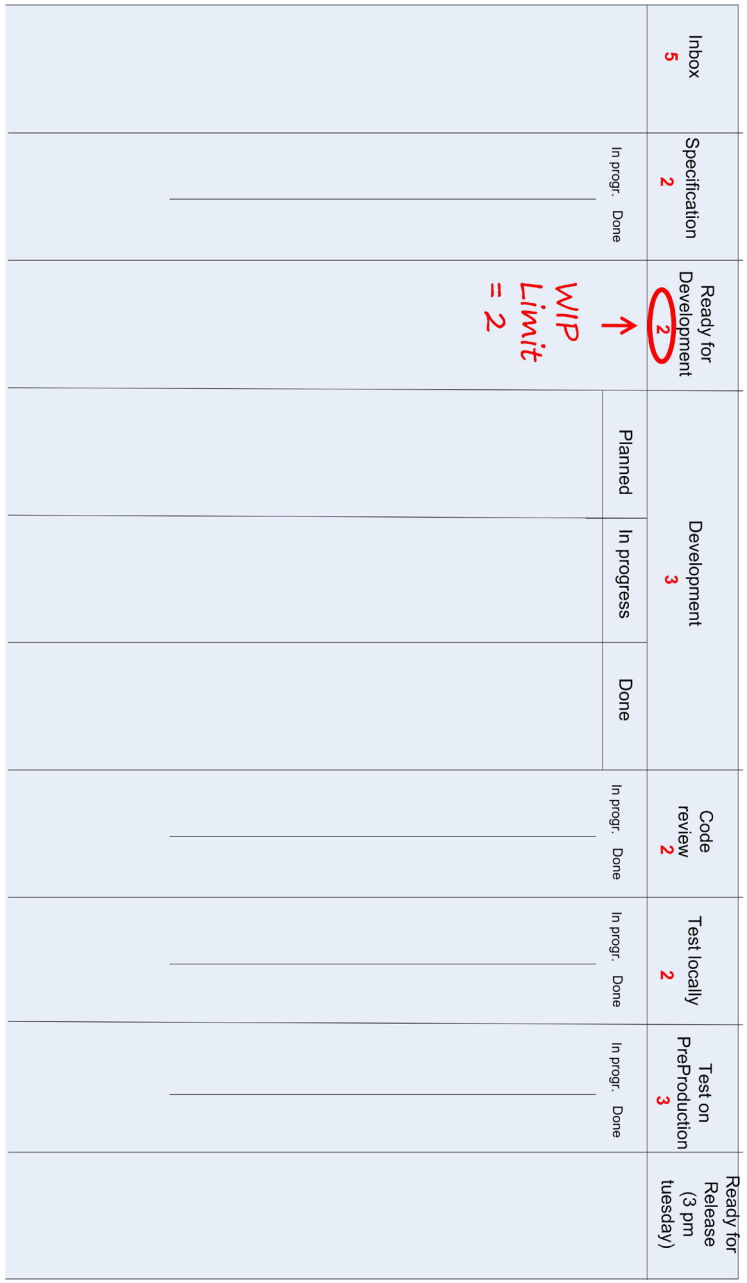

Fig. 6 WIP Limits Visualized Using Numbers in Column Headers

Finding the right WIP limits

There are many schools of thought, in terms of how tight you should set your WIP limits initially, and it is out of scope for this mini-book to cover the subject in detail. One way is to observe your system and set the limits just loose enough for your current workflow to continue unhindered. Then, identify your bottleneck and adjust one limit at a time. A more radical approach is to set your limits on activity columns tighter than you expect your system to be able to handle and buffer each stage. Then, you observe where work builds up, and gradually loosen until work flows through the system. Both require some experience, so do not expect to get it right the first time. There is no final conclusion as to which one is better, but setting the limits with your policies in mind seems to work in both circumstances.

In any circumstance, it is important to set an explicit policy of how the decision to break or change the limit will be made. To maximize learning, it is a good idea to make such decisions together as a team. This ensures that everybody gets to voice his or her opinion and understand the decision. This is not just a matter of flow but also a learning point!

Always remember that your initial limits are just best guesses, given at a place in time where you had the least amount of information available. As you gain more information about your system, limits should be adjusted continuously as you find the more optimal ways of working. If you are still working with your initial limits as well as the same stages 3 months after you started, there is a good chance that you have missed the most important step in this guide, namely the continuous improvement step we will cover in more detail later. Limits that are too tight will block the flow and will make people idle for too long or will simply be ignored without serious discussion, while limits that are too large will increase cycle time and will make work items idle for too long.

What you will quickly notice is that with WIP limits in place, your system can only work to capacity. You need to finish work to get a permission to start a new thing. While sounding trivial, it is the core concept of a Lean pull scheduling system and an incredible powerful tool on your journey to a more effective, sustainable and predictable software delivery system. A popular metaphor is to think of the system as a chain of paper clips.

As long as you are pulling it across the table, they follow a nice line, but if you push it instead, they all crumble together in a mess, each item blocking the rest (Illustrated in figure 7).

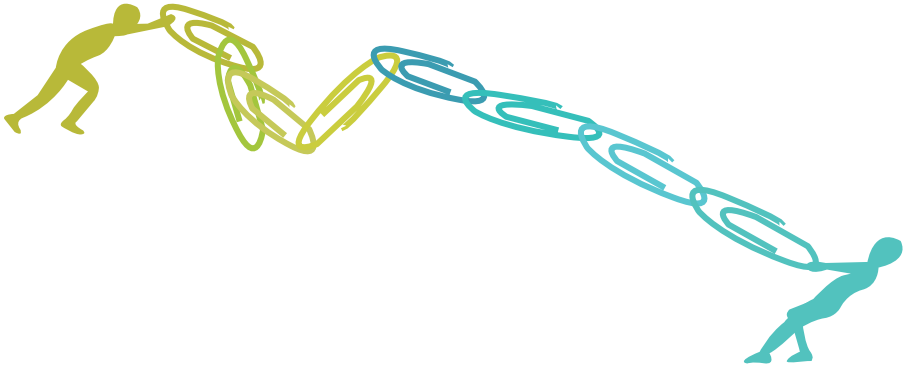

Fig. 7 Pull vs. Push

You will also quickly notice a certain pain that happens when you don't get permission to start a new thing, while you think that it is "the right thing" to do. This is a sign that you are discovering an impediment to flow – and the most important thing is not to be "Comfortably Numb" about the pain but to leverage it for improvement.

When discussing WIP, we often focus narrowly on the amount of items in progress. We should, however, not forget that size is just as important. Large items will block resources for long periods of time and will create disturbance in flow, while smaller items will flow much quicker through the system and will give us immediate feedback. Breaking items down to the minimal marketable feature set (MMF) is, however, a difficult task and it requires imagination as well as skill and experience.

Step 3 Set Up Quality Assurance Policies and Make Them Explicit

If you are familiar with Lean, you might have come across the term “Quality built in”. Why is quality so important, you might ask? Isn’t the main thing that we fix the bugs that find their way into production? The simple answer is that quality issues are much more expensive than you think. In Kanban, we focus on making policies explicit to optimize quality and consistency in our software delivery system, and we use that as a base for continuous improvement.

Understanding quality

Whether it is a user that cannot complete a task because the system lacks an intuitive interface or it is a bug that blocks the workflow, they are both quality problems and they both stress our system and generate a whole loop of waste. This is better known as “failure demand” in Lean systems and describes all activities and additional work related to the product not being designed properly in the first place. Sometimes it is acceptable, since releasing the software to get real feedback was the cheapest way to buy information. It might also be the case that finding this bug up front would have cost us a lot more time and money than fixing it afterwards. In the majority of cases, however, it is simply caused by an immature process.

So why is it so expensive? Let’s look at a couple of common scenarios in software development:

A user (let’s call him John) cannot complete his task because of a bug in our system. John writes a bug report or calls first level support to address the problem. John, however, knows little about what it takes for a developer to be able to investigate the issue or maybe the first level supporter does not know the system as well as he should. In both cases, wrong or inadequate information is given to the developer who ends up solving a different problem (which might not actually be a problem, but instead leads to another bug) or simply gives up. This continues until finally, the bug is solved. However, not only did John not complete his task, faulty information was actually saved to the database, which now needs a complicated SQL script to be reverted to a meaningful state. It is not uncommon for situations like this to occur as well as a factor of 100-1000 time and resources spent, compared to having spent the time not introducing the bug in the first place.

Moreover, quality problems cause us to task-switch and fire fight. We naturally set aside the work we are currently doing to fix a serious bug or usability problems in production, when after half a day the problem is finally fixed we cannot remember the complicated problem we were working on and have to spend an extra half an hour getting back into the context.

For these reasons, Lean puts a huge emphasis on fail proofing the delivery system (Poka Yoke). In production systems, Poka Yoke is done by using standards and checklists that must be followed when completing a task. Even photocells are used to register whether a specific screwdriver is used the correct number of times or all parts needed have been removed from the stack. This might sound like an inhumane environment to work in, but actually workers in Lean production systems do not see themselves as robots blindly following standards and checklists. They see themselves as expert operators that are constantly trying to improve the system they are working in by coming up with new ideas as to how it can be improved. In “The Elegant Solution: Toyota’s Formula for Mastering Innovation, 2006” Matthew E. May refers to this fact as the main reason Toyota still manages to IMPLEMENT one million new improvements every year.

Visualizing policies

So how does this translate to software development? Well actually, you are half way there. You have already visualized your work on the board and have put WIP limits in place. All you need to do now is add the policies you are already using to ensure quality and consistency. Doing this, your board might end up looking like the one shown in figure 8.

Note that entire stages may be QA policies and that policies also serve to ensure consistency and quality in the process itself (e.g. tracking cycle time and defect rate). All of it traces back to the third Kanban principle “making policies explicit”.

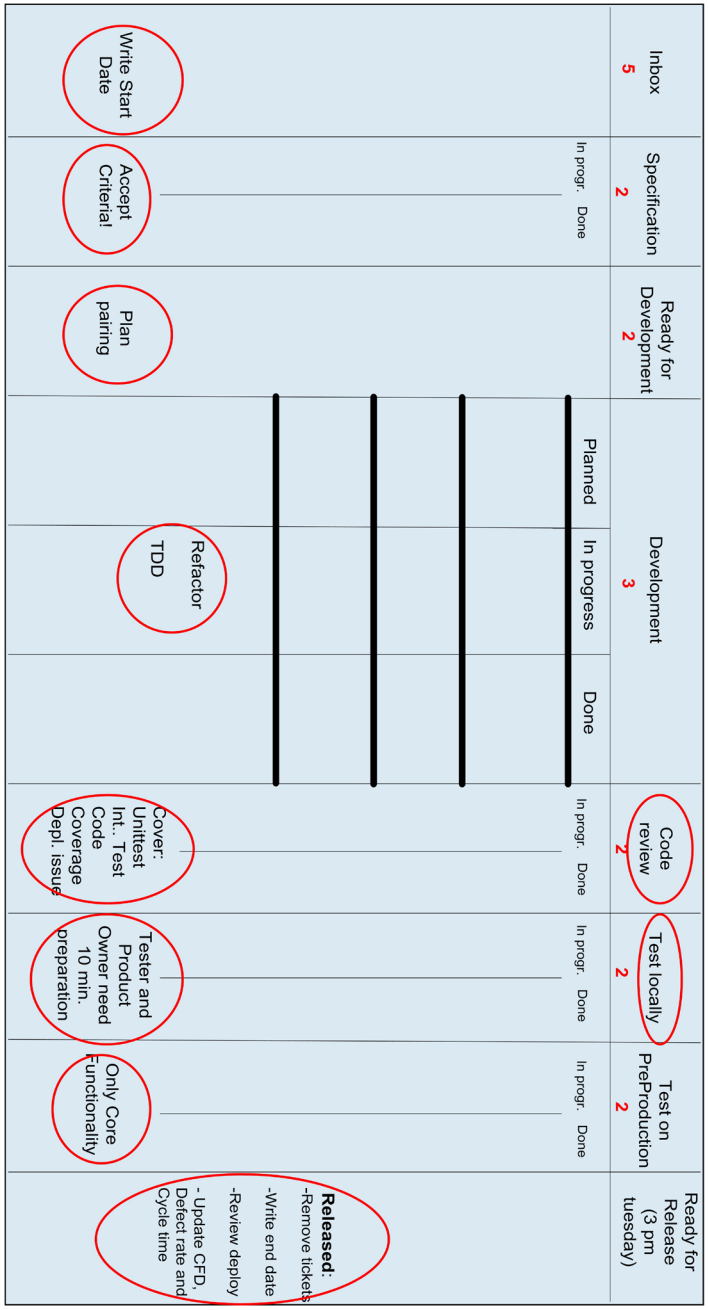

Fig. 8 Explicit Policies Visualized on Board

Many teams have taken fail proofing to extraordinary levels in recent years and have implemented systems that were unthinkable just a few years ago. The “Lean Startup Movement” has shown that it is possible to deploy software to production several times a day without down time or high defect rates. This can only work because they have put unit, integration, regression and performance tests in place that help validate the code as well as Key Performance Indicators that signal an alarm, whenever the system is not behaving as expected, and which automatically roll back to the previous versions. This makes it impossible to introduce whole classes of errors.

Always keep in mind that you should never feel that you are slaves to your policies and checklists. You are an expert knowledge worker constantly observing and trying to improve the system you are working in, not a robot. Start by adding the procedures you are using at the present time and add/remove/change them when you discover new and better ways of ensuring quality. Every bug is a chance to reflect on how it managed to get into your system.

Initially, the thought of reflecting and learning from each and every bug will seem daunting. With time, as quality improves, it will, however, become more and more natural. This is a price worth paying and is referred to as ”Zero Tolerance” in Agile Testing circles. Often, people mistakenly interpret this as “we are now allowed to spend all the time we like fail-proofing the system and all that counts is that we have zero defects”. “Zero Tolerance” does not mean that we do not expect bugs to occur ever again but to consciously reflect over each bug and strive for perfection.

When implementing Kanban, it is essential that everybody commits to adhering to the agreed policies (initially they just describe your current process) and that it takes a team decision to change them. When individuals decide to break policies on their own, without involving the team, the process quickly starts to degenerate and suddenly only ½ of the items are actually visible on the board and results are low quality and inconsistent. Remember, when you want to break a policy, it is often for good reasons and that is the perfect starting point to initiate a needed discussion. “Change them; do not break them” is a phrase I always repeat several times when doing Kanban training.

Step 4
Adjust Cadences

Once you have managed to visualize your flow, limit WIP and establish QA policies, one of the first things you should evaluate are your cadences. In a typical software delivery system, a number of activities benefit from regular cadences and finding the right one for each type is paramount to increasing flow. Note that in a Kanban system, we are not obligated to synchronise everything to the lowest common denominator. We can adjust the cadence of each activity to its own optimal level. Typical cadences we need to consider are planning (input) cadences and delivery (output) cadences. A lot of other cadences, like a review/retrospective cadence, a quality assurance cadence (if you are not a true Agile project) and regular stand-up meeting, of course, also exist; however, let us stick to planning and delivery for now.

Understanding Cadence

Finding the right delivery cadence is one of the most important things in Lean Product Development (LPD), since it helps you optimize essential feedback loops, reduce risk and optimize your delivery process. Fast feedback is the very core of Agile and Lean product development and when coaching teams, I continuously stress that work, which has yet to generate value, should be regarded as a hypothesis we need to test as quickly as possible. In his book “The Lean Startup, 2011”, Eric Ries takes this one step further and uses the concept of validated learning to measure progress. This is an interesting idea that I would suggest you take a closer look at, but which unfortunately is out of scope for this book to cover.

Though releasing every feature directly to production is the most optimal solution (given that you have a system optimized to handle this), in reality, most projects work with two delivery cadences.

- One cadence where code is deployed to a preproduction system to obtain initial feedback (internal release cadence)
- One cadence where the new version is deployed to the actual production environment (external release cadence)

Especially on “Greenfield” projects, these two cadences can be very far apart. It may take 3 months before the system is feature-complete enough

to hit production while code is being deployed and tested every day on the preproduction environment.

The important thing when choosing the right internal and external release cadence is to be aware of the economic cost of your choice. There is always a transaction cost (the cost of moving your version from one environment to another) associated with a release and there is always a holding cost associated with waiting. The balance between these two describes the optimal cadence, which is visualised in figure 9 from Don Reinertsen's book on Lean Product Development Flow (used with permission).

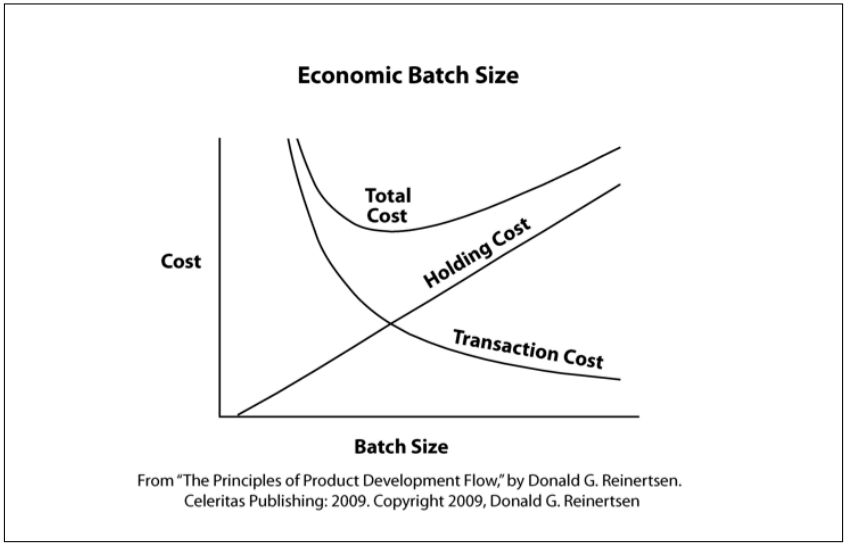

Fig. 9 Batch Size Optimization

The more features you bundle together in a release, the cheaper the cost per feature (lower transaction cost), but also a higher holding cost, since each feature will have to wait longer getting deployed, thus resulting in a loss of business value, outdated feedback, uninformed decisions and lower user involvement.

Holding costs are slightly different for external and internal releases. Since an internal release does not expose any real business value to customers, holding costs represent only outdated feedback, uninformed decisions and decreased user/business motivation, due to low involvement. Anyone having worked in a real business context will, however, recognise that these can be as detrimental as lost revenue.

As you can see from the figure, the "total cost" u-curve has a pretty flat "bottom". Therefore, it does not really matter if you hit the optimal release cadence. Being 10 or 15 percent off will still generate a good result.

Finding the right cadences

The problem is that many projects do not even consider this very carefully. Many mature Agile teams are able to release to preproduction environments with the click of a button, but still they wait a full 3 weeks before getting feedback on a given feature from users. When transaction costs can be measured in single dollars, you should strongly consider working with very small batch sizes. Sometimes this is problematic, since users are not available; however, in most cases, it has simply not been considered.

Another key consideration, which Toyota taught us, is that transaction costs are not fixed. The continuous deployment movement has shown us that it is possible to deliver reliable versions to production 50 times a day for systems handling millions of dollars. This can only be done by having a fully automated deployment procedure and a whole suite of unit, integration and regression tests, which is, of course, an investment, but is one that allows you to work in batch sizes of a few lines of code.

Adjusting the planning cadence should be done with similar considerations. When the time between planning meetings gets longer, more stuff has to be planned in one large batch. This results in more design in progress and less informed decisions, due to a longer cycle time. On the other hand, meeting everyday might prove too large an overhead and will raise transaction costs. In some cases, Kanban teams choose to plan on demand instead. This can be done by sending an email to stakeholders whenever e.g. 3 slots are empty in the input queue and by arranging a meeting or a conference call to fill them in with the highest priorities. Usually, "on demand" planning is reserved for more advanced teams and requires some prior experience handling flow. Therefore, consider carefully whether this should be your initial strategy.

Step 5
Measure Flow

Measuring progress is unfortunately one of the most misunderstood and misapplied aspects of software development we come across. Often, metrics are used to hold project managers accountable for aspects they had no control over in the first place or as fixed success criteria, established when people knew the least about the system to be developed.

Understanding Metrics

When discussing metrics, I find that one ground rule should always be remembered: “your software delivery system only has a certain capacity”. If you try to press your system beyond its capacity, it will lead to lower quality, unsustainable pace, higher maintenance costs, or all of the above. But still, time and time again, we see project managers almost bragging that they have made their teams work overtime for 3 months or that by some heroic effort, they have fixed things at the last moment when everything was total chaos. Though we should celebrate great achievements, software development projects do not need fixers; they need people that are able to deliver with transparency and a healthy sustainable pace. Everything else is simply too expensive. I like Kent Beck’s statement that “if you have a problem that requires more than one week of overtime, you have a problem that should not be fixed by working overtime anyway”.

You may of course increase your capacity over time by hiring more people (beware of doing that for short-term results) or by optimizing your process. Another good ground rule to consider here is that “your system never has more capacity than it has PROVEN to be able to deliver”. Following this simple rule will also keep you from managing projects by the anti pattern of “wishful thinking and other people’s successes”. Starting on a Greenfield project, your capacity and capability will of course be informed guesses from previous performance, and the key here is to track progress from the beginning to validate those assumptions. There will be more on that topic in step 8.

So, think of your plan as a tool for alignment, not a success criterion, and measure your flow to determine whether you are still aligned. What we want is our software delivery system to be stable and predictable, so that we can make informed decisions about deadlines, dependencies, staffing, scope and budget.

What to measure?

So, how do you measure flow? There are dozens of ways to do this and the main thing to consider is always “will I act on this piece of information”. If you are not going to change anything based on a chosen metric, chances are that you shouldn’t be measuring it at all. If you have no idea where to start, I suggest starting with the following four: Cumulative Flow Diagram, Cycle Time, Defect Rate and Blocked Items.

Cumulative flow diagrams (CFD)

Cumulative flow diagrams seem to be replacing burn down charts for more mature Agile teams and organizations for good reasons. They are easy/easier to update and give you better insight into the system’s status. For those unfamiliar with the concept of CFDs, they simply display the current amount of work in your system for each stage over time. While this may sound simplistic, it provides you with the same kind of information as the traditional burn down chart, plus a lot more. Figure 10 shows an example of a CFD.

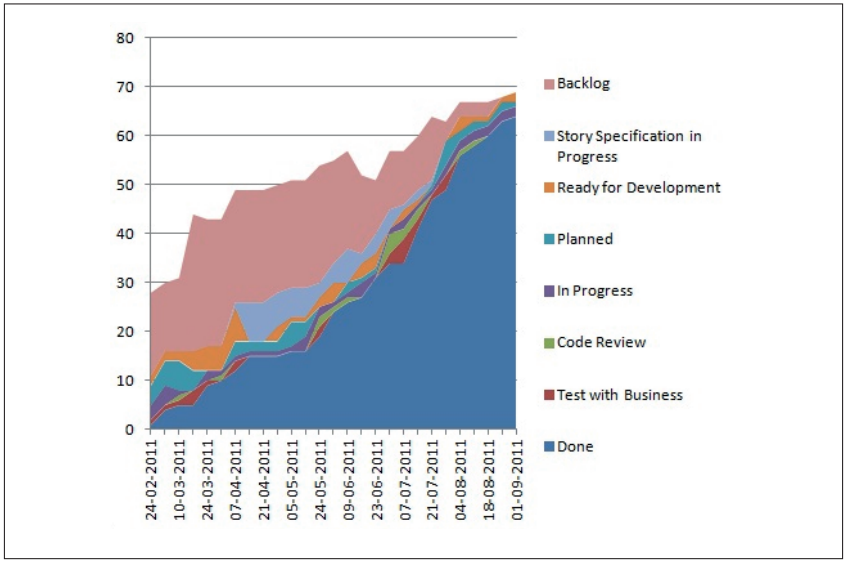

Fig. 10 Cumulative Flow Diagram Example

Reading the CFD

The gradient of the “done” area describes your velocity over time, while the space between this line and the “backlog” line may be defined as WIP.

- If the width of a part of the WIP area increases, it could be a sign that a bottleneck is occurring.
- If the gradient of the “backlog” area is steeper than the gradient of the “done” area, it is a clear sign that you are adding more work to your system than your current capacity.
- Projecting where the gradients of “backlog” and “done” cross is your current best guess of a final release date.
- Average Cycle Time and Quantity in the queue can also be established from the diagram.

Learning to read a CFD is easy and figure 11 from Don Reinertsen’s book (used with permission) gives an excellent visual representation. The black area equals WIP.

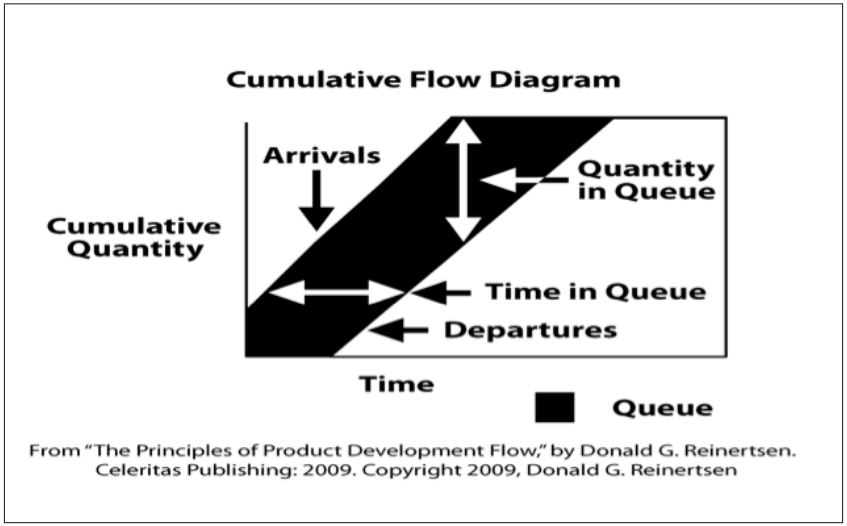

Fig. 11 How to Read a Cumulative Flow Diagram

Cycle time

Though your Cumulative Flow Diagram will tell you the average cycle time, tracking individual cycle times can be very helpful in terms of predictability.

Averages can be misleading and a visual representation will give you detailed information about the reliability of your system as well as the opportunity to meet customer demands more accurately (something we will cover in more detail in step 9).

Tracking the Cycle Time is even easier than updating the CFD. All you have to do is register the date work started on an item (remember to make this policy explicit as well). When work has finished, you plot the number of days it took to complete and your diagram should look something like the one shown in figure 12. Since each "step" on the x-axis simply represents a completed work item, teams often choose to leave it without "unit".

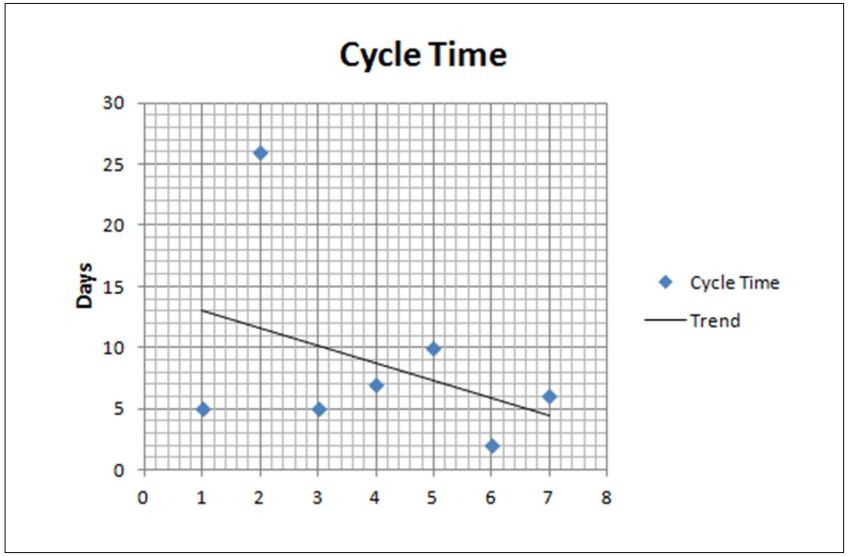

Fig. 12 Cycle Time Diagram Example

Though simple, a cycle time diagram tells you a lot about how your system is working:

- Do you have a high level of consistency or are the numbers far apart?
- Is the trend going in the right direction?
- A chance to investigate outliers (positive and negative).
- The consequence of decisions (large tasks, fire fighting, quality issues…).

If 90 percent of work items take under a week, you might want to tell your customer that they can expect that 9 out of 10 times, work will be completed within a week.

You should, however, remember that Cycle Time is a lagging indicator. This means that we will only see problems after they have occurred, where naturally it is too late to do anything about it. Therefore, we have to use cycle time diagrams, together with e.g. a CFD, to be able to act proactively.

Defect rate

As previously mentioned, quality issues are incredibly expensive and you therefore want to keep them under a watchful eye. Tracking the defect rate and the total number of bugs in your system is an easy way of making sure that quality problems do not get out of hand.

Surprisingly, few organizations use defect rates as a KPI (Key Performance Indicator), despite the fact that it does tell you a lot about the status of your project:

- Why is the number of new defects increasing? Did you relax some QA policies?
- How did the high level of bugs in week 20 affect the cycle time?
- What was the impact on the cumulative flow diagram when the number of bugs increased?

As usual, always be careful not to make too many conclusions based on individual data sets. A bad week might just be a coincidence. Look at trends to see if you are moving in the right direction. Figure 13 shows an example of a defect rate diagram.

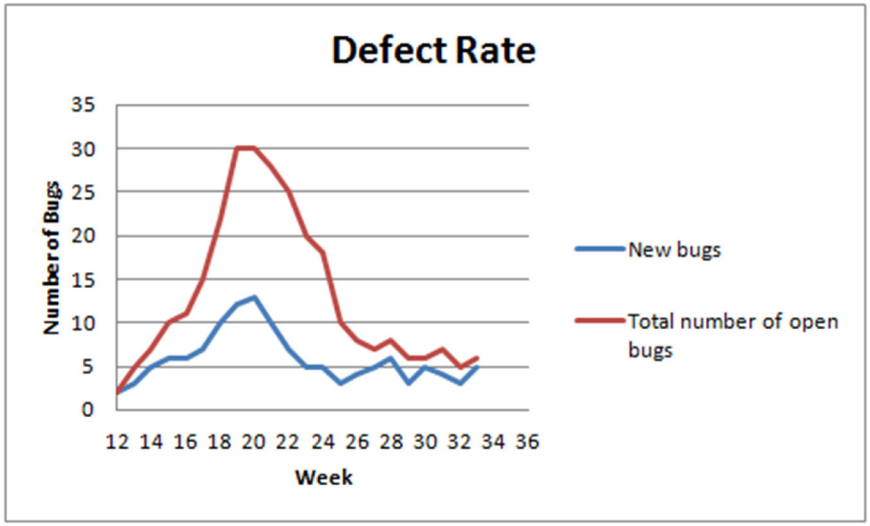

Fig. 13 Defect Rate Diagram Example

Keeping the total number of bugs between 0 and 20 is a good policy for most projects. Once the list gets bigger, it becomes hard to administer and you have to spend time checking for double entries, outdated issues and things that have already been fixed. People also seem to get nervous and demand more reports and tracking once the list approaches 50 or 100; before you know it, there are bug management boards and weekly bug meetings stealing your valuable time. Even cosmetic bugs require attention and take time to administer, so don't fall into the trap of allowing 50 of them either. Often people will try to categorize their way out of trouble, by simply changing the severity of the bugs they have registered. It is however a very short sighted solution due to the reasons stated above.

Blocked Items

By now, I hope that you are convinced that flow is important for our systems' ability to act predictable and for the individual processes to operate effectively. Most people working in both Agile and non-Agile contexts will have experienced items being blocked for longer or shorter periods of time and for various reasons. Though this will show up on our CFD and eventually our Cycle Time diagram (if the item makes it through the system), most teams find it beneficial to explicitly and visually track the team's ability to handle and fix issues blocking one or more features in the system. Some companies even use this as the leading Key Performance Indicator (KPI), since they recognize that blocked items have serious long-term effects on the systems and that a team's ability to quickly solve issues says a lot about the team's performance and effectiveness. Blocked items should always be visible on the board, and tracking the status over time is usually a good way of knowing whether the team is moving in the right direction. Figure 14 shows an example of a blocked items diagram.

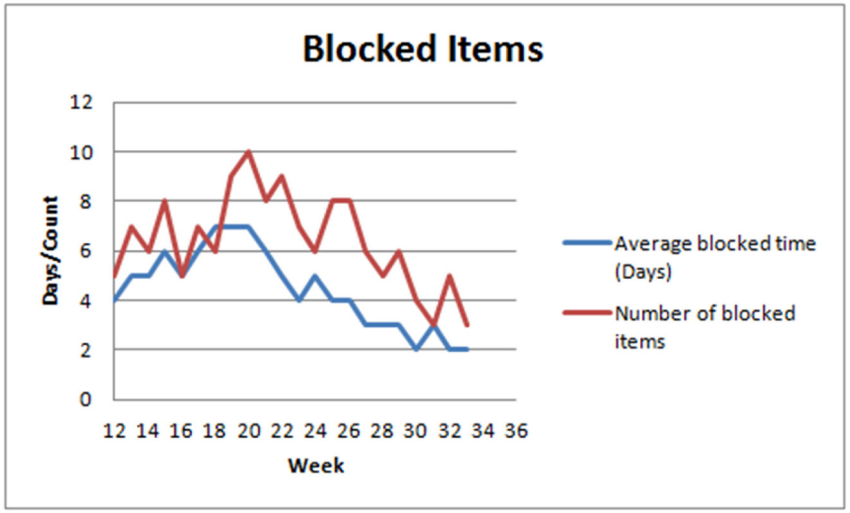

Fig. 14 Blocked Items Diagram Example

The standard way of visualizing blocked items on the board is simply to attach a pink sticker to a particular feature, with the blocking issue and the date it became blocked written on it. Figure 15 shows a board where a pink sticker marks a blocked item.

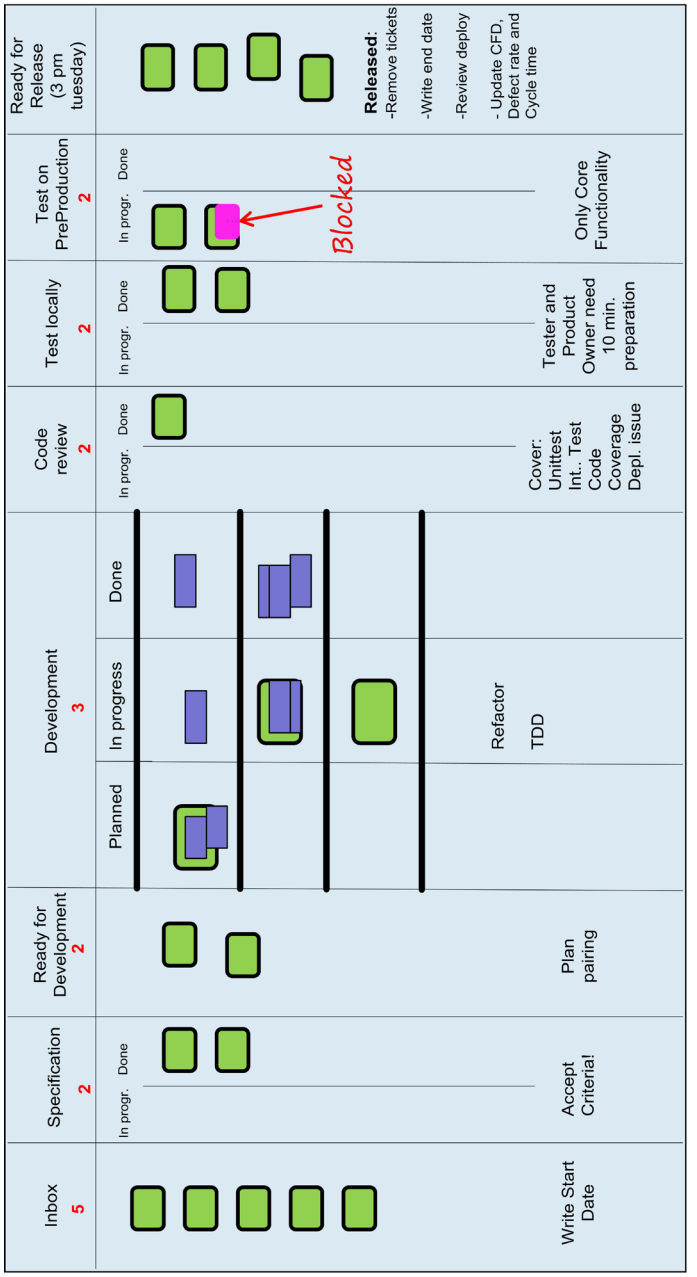

Fig. 15 Blocked Item Visualized with Pink Sticker

Try to avoid having a particular place on the board for blocked items. Since this place is not part of the actual workflow, there is a tendency that people grow numb to these issues and they end up having their own little corner where they rarely get attention (before someone turns up yelling and screaming, wanting to know why it was not finished two months ago). It also seems to generate a behaviour where people become less interested in resolving them and more interested in stating that some external party is currently responsible for getting it solved. This might be true, but it still represents a problem that you have invested time and resources in something that is not moving and has not yet generated a single dollar of revenue. It is, however, just my personal opinion. You should choose the strategy that fits your own context.

Four diagrams, next to the board, is often the limit as to how much information most people are able to process before they simply "drown" in it and start to care less. It is much better to use one metric actively than having four you rarely pay attention to. In any case, it is important that metrics are posted visibly, and keeping them hidden on a separate sheet in an electronic system will rarely get much attention. Figure 16 shows the four diagrams, covered in this chapter, on top of the board.

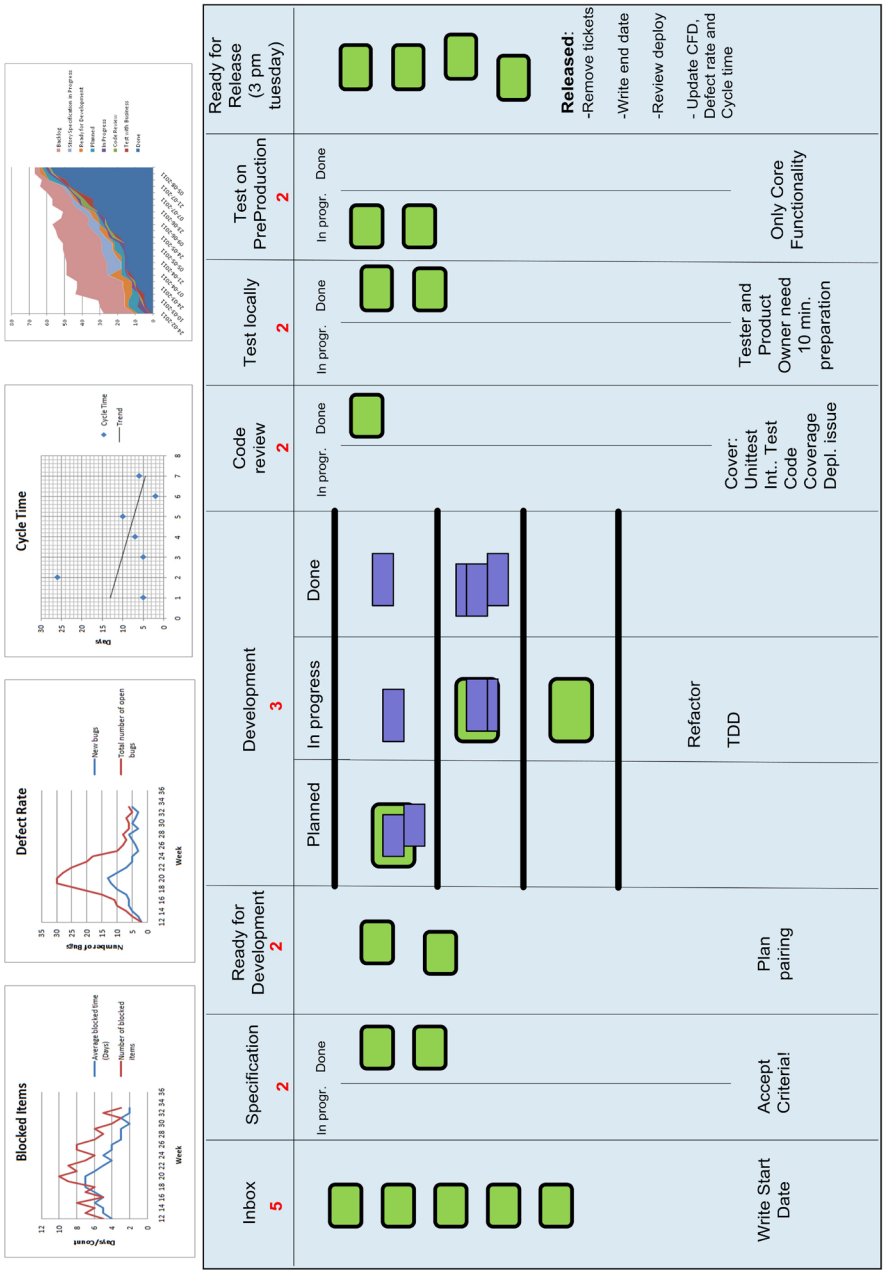

Fig. 16 Flow Metrics Visualized on top of Board

Step 6
Prioritize

It may come as a surprise to someone that we are down to step number 6 before starting to deal with pulling things in the right order. The reason, however, is quite clear. In his book “Kanban”, David Anderson states that if you don’t have a working software delivery system, which is able to deliver reliably and with quality, your prioritization matters little. In this case, you should probably spend your time fixing the problem of not being able to deliver first. This is of course context-dependent and in Greenfield projects, you might want to consider this earlier on. In any case, there is no reason to stop your current way of prioritizing work, so keep doing that and consider using the following strategies when you are ready to use a more Lean way of approaching prioritization.

So, how do we prioritise our work the best way possible? In step 7, we will look at how different types of work should be handled differently, but for now, we will stick to the prioritization of one type of work, e.g. “user stories”.

Cost of Delay (COD)

The default principle is Cost of Delay (COD), and Don Reinertsen again is by large responsible for introducing this principle to IT. COD describes the revenue or expected cost saving lost by choosing NOT to work on a given item. Your highest priority should be the item with the highest COD. In reality, the COD will often be weighted by Cost of Implementation (COI), deadlines, time and other factors. It is out of scope for this book to explain the full concept of COD. For now, all you need to do is wrap your head around the concept of lost opportunities and that every time you choose to work on something, you are choosing to block something else. Calculating the exact COD is almost never possible in IT, so we will often have to do with our current best guess based on the available data. A good guess, however, is much better than no guess at all and learning to place economic value on your work is a maturing exercise for all projects and organizations.

Visualizing Priority

To make sure that we pick the right item to work on, our input queue should always be prioritised and new work pulled from the top. This rule applies no matter whether you are working with Scrum, planning a batch of work for the next sprint, or with the flow-based approach continuously pulling the highest priority when a work permit exists. Figure 17 shows an example.

Other important prioritization factors, which should also be included in the final rank, include:

- Risk and uncertainty. Buy information early for high-risk and high-impact decisions
- Bare necessities: Project infrastructure, etc.
- Balance size: Mix story size to keep a steady flow.
- Balance story types: Mix functional/non-functional stories to ensure a steady flow of value.
- Dependencies: Handle dependencies proactively so that work does not get stuck.

If you are working with a traditional backlog or a waterfall-like requirement specification containing 50+ items, people naturally start to question how much of it to visualize on the board. There is no general rule; some teams find it helpful visualizing the entire input queue, while others keep the list in a separate place/tool and gradually pull the e.g. 5 most important items on the board. I like to use an iceberg as a metaphor for describing this policy. Only the top of the iceberg is visible above the water, but if you remove it (pull the items into your system), ice from underneath will emerge to form a new top. Keeping an explicit WIP limit on the input queue is, however, a very good idea to keep it from spinning out of control. I usually compare a backlog to an unused top floor of a house. If you put everything up there that you are reluctant to throw out, you have very little chance of finding the few things you actually need when you need them. Therefore, keep the backlog clean and make sure it is not growing out of control.

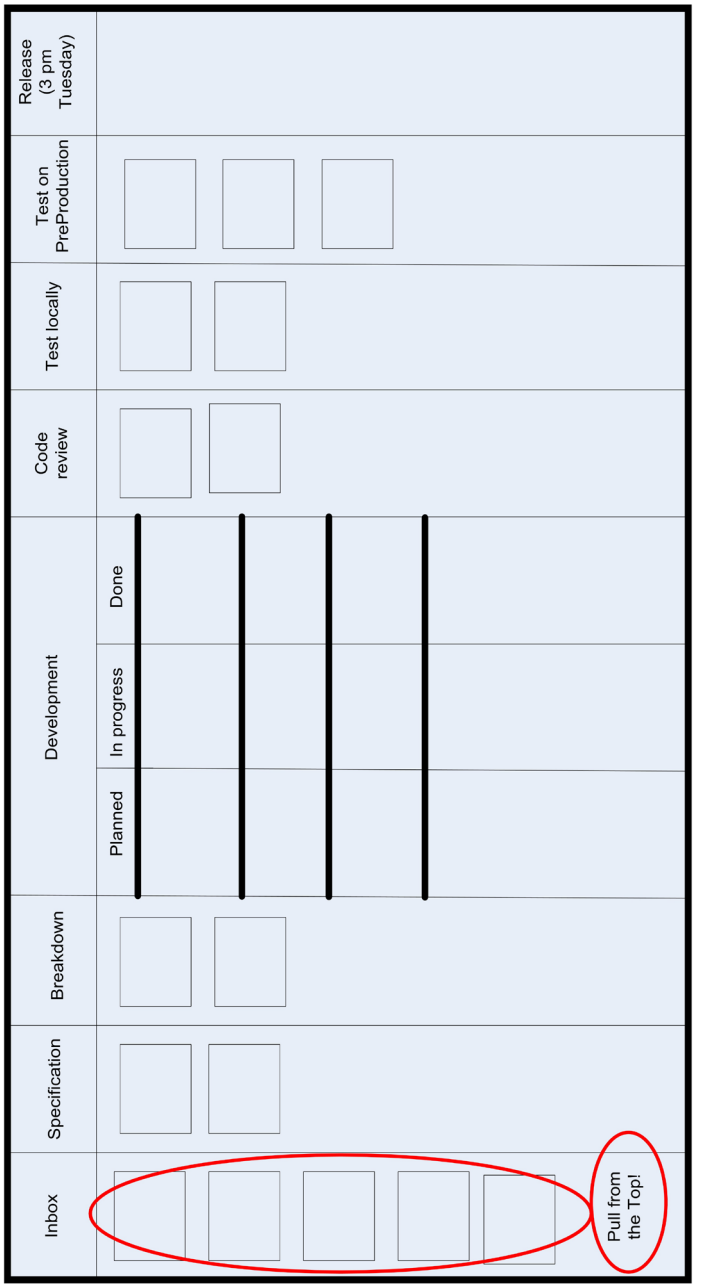

Fig. 17 Prioritization Policies Visualized on the Board

Step 7
Identify Classes of Service

Not everyone is created equal and the same goes for the way we deal with different types of work in software development. Few would question that an issue resulting in 10,000 users being unable to access the system and costing $100,000 in revenue per hour deserves special treatment, compared to a feature under development. But how do we make sure that we choose the most reasonable way of processing these different types of work?

In a Kanban system, the way of doing it is referred to as “Classes of Service”, which simply mean that we will treat things differently according to their specific characteristics.

So how do we approach establishing classes of service?

Types of work

Different types of work exist in all software delivery systems and identifying these is often a good starting point. Individual work types will differ from system to system, but almost all have some element of requirement, e.g. User Stores or Use Cases and defects/bugs. These may again be divided into categories of functional and non-functional user stories, as well as blocking, critical and cosmetic bugs.

Typical types of work include:

- User Stories (Small, Medium, Large)
- Bugs (Cosmetic, Critical, Blocker)
- Manual Reports
- Textual Edits
- Support Tasks
- Installation

Define Classes of Service

Once you have defined your different work types, the next step is to consider how you will handle these different work types in your system. Each way of handling work types is a Class of Service. The best way to explain this is by showing an example. In the following, we have defined 4 classes of service.

Standard Class

- Extra cost: 0
- Work types: Cosmetic Bugs, User stories
- Special treatment: None

Priority Class

- Extra cost: $500
- Work Types: Critical bugs, High priority user stories.
- Special treatment: Takes priority at each stage.

Fixed Deadline Class

- Extra cost: $ 0-2000
- Work Types: User Stories
- Special treatment: Takes priority at each stage if deadline is deemed unsafe. Otherwise, it is treated as a standard class. Emergency de ployed if necessary.

Expedite Class

- Extra cost: $3000-5000
- Work Types: Blocker Bug
- Special treatment: Break WIP limits, stop existing WIP, emergency deploy

Special treatment defines how this class differs from a standard work item when introduced into our software delivery system. There is always a cost associated with giving things special treatment. By measuring flow, you should be able to make a qualified guess. What is the effect of giving something specific treatment? How much longer will it take for the rest of the items to get through (given you can estimate cost of delay)? How much time/extra time will be used in total, due to task switching, and will you have to spend extra time deploying, etc.? Initially, it will, however, always be a best guess on the average cost and some teams choose to collect a few weeks of data first. It is, however, an extremely powerful tool and to some people, it is a huge surprise to discover that expediting is not free. This will naturally cause everybody to evaluate whether it is worth doing it. Once you start to measure your flow, you will be able to make more informed guesses about the cost of special treatment. You can see the effect expedites have on the cycle time diagram as well as how an emergency deploy blocks flow and consumes resources.

Often, it is a good idea to set a fixed limit on the number of non-standard classes in our system. A good rule to consider is: *"If everything is an expedite you have got no expedites at all".*

Visualizing Classes of Service

Classes of Service can be visualized in a number of different ways. Two of the most popular ways of displaying it is either using color codes (figure 18) or swim lanes, as shown in figure 19 (or a combination of both).

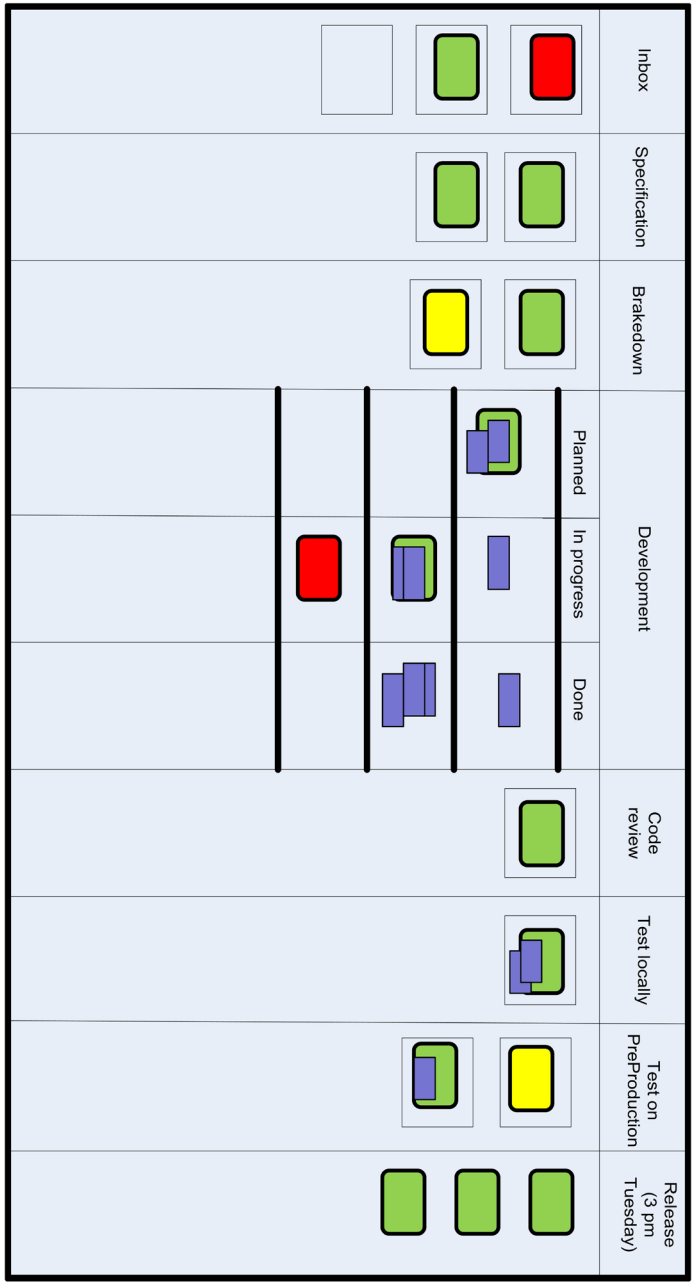

Fig. 18 Classes of Service Visualized Using Color Coding

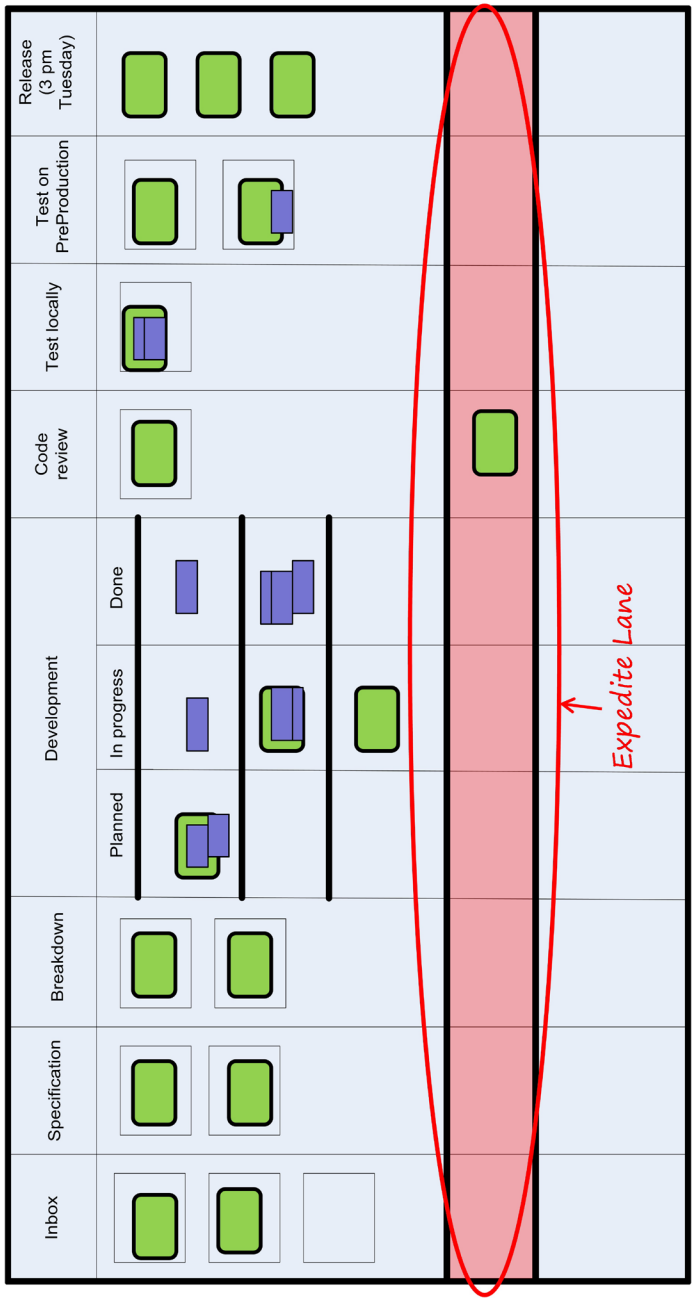

Fig. 19 Classes of Service Visualized Using Swim Lanes

Using classes of service gives you the opportunity to handle each item in a rational way, according to its economic impact, instead of resolving to panic and fire fighting. It also means we can make different promises to our customers, depending on the class of service we are handling. This is a topic we will cover in more detail in step 9.

Step 8
Manage Flow

By now, you are already showing signs of operating in a highly mature Agile environment. You have visualized your entire workflow, limited WIP, set up QA policies and have started tracking your flow. The next thing is to learn to read your system and to take appropriate action when you see an improvement opportunity.

Decision filters

In general, I find it useful to use David Anderson's Agile and Lean decision filters to guide our actions.

Agile Decision filter

- Are we making progress with imperfect information?
- Are we encouraging a high trust culture?
- Are we treating WIP as a liability rather than an asset?

Lean Decision filter

- Value trumps flow
- Flow trumps waste elimination
- Eliminate waste to improve efficiency

While the first two points in the Agile decision filter are of a more broad character, the rest can be used to make better and more informed decisions when dealing with challenges and difficult decisions.

What the Lean decision filter simply states is that value is more important than flow, so be careful when trading value for a better cycle time. This is actually a common problem in Agile projects where business value (and sometimes also quality) is often sacrificed to get more things done.

In one of the worst cases I have seen, 3 teams working on the same product had been forced to work overtime for the last 3 months. When I asked them "why", the only answer I would get from team members, product owners and the program manager was that they had to complete a long list of features and correct a number of bugs. When I asked what business goal they were trying to reach, most would just stare at me. When I finally convinced them that they had to agree with the customer on a shared goal/vision for the release, instead of just working blindly on a long list, I felt we were getting somewhere. After 1 week of hard work and negotiation, the team of product owners and the program manager presented the vision to the teams. Most were happy but one team member raised his hand and said "are you aware that we have not got a single item on our board that will move us one step closer to that vision?". Though they tried to give him an answer with a straight face, the embarrassment was showing clearly. The team probably had at least 10 items in various stages on their board and if not, a single one of them moved the project closer to the new vision; having worked overtime for such a long time seemed very much like a waste of time and resources. This was a valuable lesson; you might have the best Kanban system in the world and great flow across the entire value chain, but if you are missing core feedback loops on your value hypothesis, that might just mean you are producing waste faster.

Flow, on the other hand, is still more important than waste elimination, so be careful when trading flow for e.g. increased capacity utilization - an issue we will cover in more detail in the next section.

When you have managed to optimize for value and flow, you are ready to look at waste elimination; however, make sure you first and foremost watch the product before the people.

With the Agile and Lean decision filters in mind, let us try to look at some core concepts for managing flow in our software delivery system.

Optimize flow, not utilization

When you look for improvement opportunities, try to avoid thinking, in terms of utilization. Look for opportunities to increase the flow of work items through your system by asking the following questions:

- Are you working with the right WIP limits?
- Can you find a way of making the size of user stories smaller?
- Is there a way to identify features that explode in size, before they are introduced into your system and end up blocking capacity for long periods of time?
- Can you level out the size of user stories to create a more continuous flow?
- Can you train for flexibility to avoid silos and easier relieve bottlenecks?
- Have you got adequate buffers in place to handle variation?
- Are you looking at optimizing the whole and not individual stages?

Optimizing flow instead of utilization is close to the very core of Lean. American car manufacturers used to measure and reward individuals according to how many e.g. car doors they could produce, even if those car doors were just stock piled in a storage building somewhere. This made the individual machines work incredibly fast, but slowed down the overall production, since large storages made it hard to locate parts, move parts around, and vast amounts of money were tied up in the inventory. The Toyota production system totally changed the game and showed that by focusing on the end product and matching the individual machines' speeds to the ration of cars coming off the production line (tact time), it gave an economic advantage that could not be disputed.

The key is to always think in terms of the flow of the end product and try not to focus on how you can make an individual or an individual step go faster. Unfortunately, many managers are still much more focused on getting people to work faster than the quality and flow of the product. Always remember to watch the product, not the people!

I recently discussed the issue with a client, where everybody was extremely focused on utilization and asked the question: "What do you do to make sure you are always busy". One replied "I have an assignment I can always work on when there is nothing else to do". I asked him how long he had been working on it and what the value was. His reply was quite interesting: "I have been working on it for a year; it hasn't been released yet so it has not created any value so far. One of his colleagues ask him when he thought it would be released: "Well, I have to admit that due to recent changes in our product portfolio, it will probably end up being killed in a week or two". The rest of the group laughed out loud and openly admitted that this was not an uncommon scenario in the company.

Relieve bottlenecks

The Theory of Constraints (TOC) teaches that there will always be one bottleneck within a given system, thus limiting production flow. Though TOC brings a simplified view to flow and bottleneck handling, it helps us to understand the importance of looking at the system as a whole and to focus our efforts where they bring most value. Using a visual Kanban pull system, bottlenecks are easy to identify, as you will see work piling up in upstream processes and the workflow being drained in downstream processes. The immediate reaction is often to add more capacity, but often there are other and more effective ways of handling bottlenecks. People simply do not scale the same way machines do and the increased capacity, in terms of the number of people, is often eaten up by the increased coordination overhead and training. Remember Brooks' Law: "Adding manpower to a late software project makes it later."

Instead, try to look for opportunities to protect the bottleneck from unnecessary work. In one project, we found the PO team to be the bottleneck. When analyzing their work, it became apparent that much of their time was consumed by bug investigation and getting back to users who had not received the necessary education to work in the system. This task could easily be undertaken by members of the development team and the result became that developers would rotate the task of doing this.

The longer-term perspective could be to find out how this came to be, and either improve the workflows in the system to make them more intuitive or help users to get a better introduction. Removing non-value adding work is by far the most effective way to relieve a bottleneck.

A third way could be to investigate whether the PO team has blocked work items consuming capacity. In this case, the team should "swarm" on these to get them fixed as soon as possible

More ways of dealing with bottlenecks can be found in David J. Anderson's book "Kanban", mentioned at the beginning of this book.

Introduce buffers

If you know that there is a bottleneck in your system, it is also good to introduce an appropriate buffer in front of it to make sure that the bottleneck is rarely drained. If, for example, your bottleneck is "Development", a buffer stage with items "Ready for Development" could be added. Choosing the right size takes some experience. It is ok for the buffer to be emptied once in a while, but if it happens every 2 weeks, you should probably choose a larger one or evaluate if this is indeed still a bottleneck (could be upstream since the buffer was drained continuously).

Release planning

Despite being called "Product Development", most teams doing actual software development (all that I have worked on) live under the "Project" constraints of budget, time and scope. Thinking only in terms of flow is therefore often a naive approach, since steering group committees expect you to be able to answer the questions: Are we on time? Are we on budget? Will you deliver the agreed scope?

To be able to handle this situation in a sensible way, you need to do two things:

Firstly, you need to agree that since you cannot fix all three, scope will remain flexible. Moving a deadline is hard and often results in a vast amount of time spent reorganising, coordinating and communicating. Increasing the budget often means adding more people and as previously mentioned, this is seldom a good tool to reach an upcoming deadline. Adding more-people is a strategic move for the long-term perspective. When increasing the budget, it does not mean more people; it means making the people you have got work longer hours. It can be a tool if you have only got one week to go, but in all other situations, you should refer to the previous statement by Kent Beck that "if you have got a problem that cannot be fixed by working overtime for a week, you have a problem that cannot be fixed by working overtime anyway". Too often, overtime is used as the project manager's desperate tool to show that "I am doing something". He knows it will not fix the problem but uses it to show some kind of action. In some rare cases, moving the deadline is the right solution, but for the reasons mentioned above, keeping a flexible scope is a good rule of thumb.

Secondly, you need to understand a flexible scope. When people hear the term "flexible scope", they often interpret it as a laissez-faire approach to software development, which really means "do whatever". However, working with flexible scope requires discipline as well as the ability to track progress accurately to ensure that you are making informed decisions in a constantly changing environment. This is a key factor in getting the best possible ROI.

You know that your original estimates, in terms of complexity, cost, and business value, were made at the point in time when you had the least amount of information available. Still, deadlines and budget were based on these assumptions and it is therefore essential to track your progress to make sure that the project is still feasible. There are many ways of doing it, but since we have our CFD in place, why not use it for this purpose as well? Most project budgets are done by release, so let us look at an example of that (doing it, for multiple releases, only requires minor adjustments). By having a budget, deadline and initial scope in place, we simply draw our expected velocity on the CFD and track our progress according to that.

Remember that most projects follow a S-Curve and that deviations, as a general rule, indicate that you now have more information available than you did before and therefore are able to make more informed decisions (could be to kill the project early).

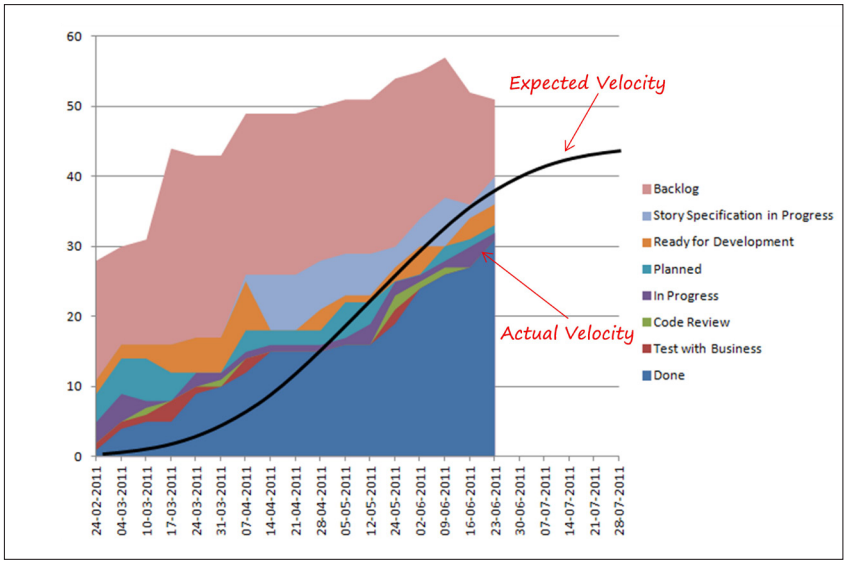

Fig. 20 Release Plans Visualized on the Cumulative Flow Diagram

Figure 20 shows an example of a S-Curve on top of a CFD. As you can see, the backlog was expected to increase by around 40 percent (from experience), from 28 to 40 points, but at one point had grown to more than double the size (57). While starting out with a higher than expected velocity, velocity dropped half way through the release (in this case because of quality problems).

Experiment

Managing flow also means trying to continuously improve it (covered in more detail in step 10). Many projects do this blindfolded in the sense that they have no means of telling whether the things they changed were a success or failure.

Unfortunately, most Agile projects fall into this category. Retrospectives are used to set up experiments, but following up (if done at all) only includes whether it was carried through or not. There will of course always be a high level of uncertainty, but more often than you think, the simple metrics introduced in this book provide a clear visual indication of whether it worked or not. As you might recall, the Deming circle includes Plan, Do, Check, Act, because it is necessary for us to be able to make informed decisions moving forward.

For example, if you decide to include testers in the development team to do more upfront testing. You should expect to see a drop in defect rates within a reasonable timeframe and be able to work with a lower WIP limit for “Test”.

Managing flow is all about reading your software system to make the best possible decisions with the information available in the pursuit of the highest ROI.

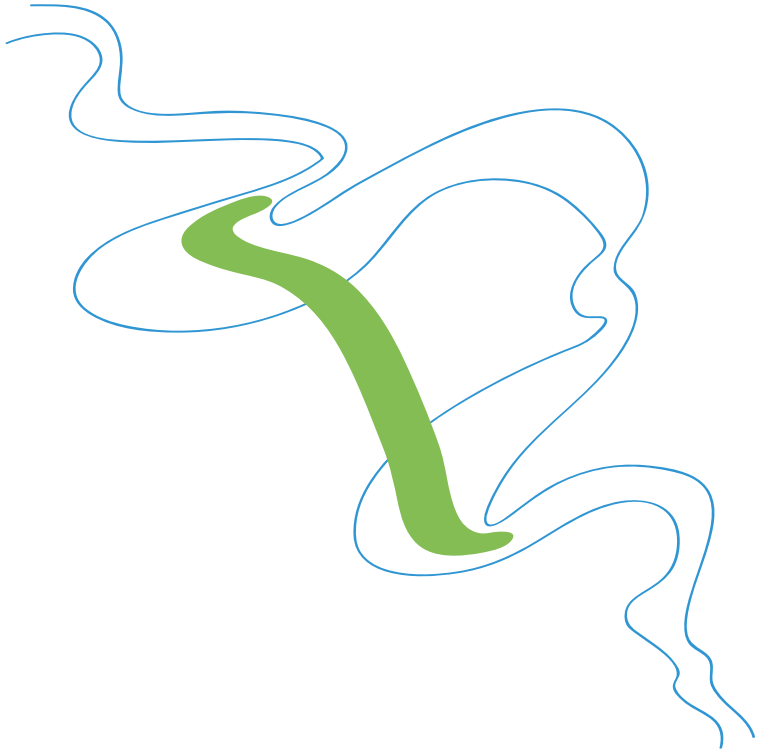

Fig. 21 Relieve Bottlenecks to Improve Flow

Step 9 Establish Service Level Agreements (SLA)

You are now well on your way towards establishing a more effective and reliable system for software delivery, so now it is time to show your results to the outside world. Having a stable pull system in place and using a simple set of metrics, to track the system's performance, will make it possible for you to establish SLAs that you actually meet. This will help you keep the system in place and avoid the traditional revert to fire fighting and chaos once the Kanban initiative is no longer new and shining. So how does this work?

Establishing the right Service Level Agreements

While traditional Agile approaches, like Scrum, put high value on predictability in terms of Sprint commitment, a Kanban system works on the belief that you will gain predictability by having a software delivery system that works in a predictable way. There is a subtle difference between these two ways of approaching predictability, which should not be underestimated. One is based on a plan-driven approach, while the other is flow-based.

If you treat your different classes of service the same way every time and measure the consequence of your improvement efforts, chances are that cycle time, quality and cost will only improve over time. This gives you the possibility of sharing this data with your customers. The previously mentioned classes of service might therefore get an SLA looking something like this:

Standard Class

- SLA:
  - o Mean: 15 days
  - o 90 percent within: 21 days
  - o All within: 30 days

Expedite Class

- SLA:
  - o Mean: 2 days
  - o 90 percent within: 3 days
  - o All within: 4 days

Fixed Deadline Class

- SLA:
  - o 98 percent within deadline

Priority Class

- SLA:
  - o Mean: 8 days
  - o 90 percent within: 13 days
  - o All within: 18 days

The key here is that we know these numbers, not because of qualified guesses but because we have been tracking our system's performance and have collected the necessary data. If a demand arises for us to provide even more detailed information, we can easily adjust our metrics accordingly. If, for example, it turns out that our standard class work items differ a great deal in size, we might want to add the following details to show our customers the direct effect.

Standard Class

- SLA 200-300 story points (Large):
  - o Mean: 21 days
  - o 90 percent within: 25 days
  - o All within: 30 days

- SLA 100-200 story points (Medium):
  - o Mean: 13 days
  - o 90 percent within: 18 days
  - o All within: 25 days

- SLA 10-100 story points (Small):
  - o Mean: 10 days
  - o 90 percent within: 14 days
  - o All within: 18 days

For many customers, this information is highly valuable in terms of prioritization and in experiencing that these numbers hold true, this gives an amount of trust and collaboration far beyond what most have experienced in prior projects. Also, this will show the direct benefit of breaking work down into smaller sizes, both in terms of risk, cost and cycle time. To make sure everybody is aware of the current SLAs and Classes of Service, most teams find it useful to post them next to the board, as shown in figure 22.

Fig. 22 Classes of Service Policies Posted Next to the Board

Step 10 Focus on Continuous Improvement

Keeping a constant cycle of improvement going is arguably the hardest and most important element when implementing Kanban. As previously mentioned, Kanban is a method for driving evolutionary change, and the good news is that having gone through the previous steps will make this process of continuous improvement a lot easier.

The extreme amount of information radiation created through the visualization of workflow, explicit policies and SLAs for each class of service has proved to foster an ongoing dialogue about improvement opportunities far beyond those seen in traditional software projects. Every day you are forced to make explicit decisions about how best to handle work in your system. While being a good base for continuous improvement, this fact also seems to provoke a deeper understanding of Agile and Lean concepts for those working in a Kanban system and is therefore less likely to revert to former processes and anti patterns.

Since we collect real data, Kanban also gives us the opportunity to perform and validate our experiments in a more scientific way than traditional Agile projects. Initiatives to improve cycle time should result in an actual measurable effect. This makes continuous improvement in a Kanban system much more reliable and since we can see and measure the effect of our change initiatives, we are much more likely to keep raising the bar.

For some people, working with Kanban is the first time they start to see the software delivery system as a whole. This gives an immense insight into other people's work, how they depend on you, and vice versa. This also means that opportunities for optimizing more than just individual silos arise from the ongoing discussions between the involved groups of people. If you happen to see a tester, PO and a developer discussing flow improvements next to the Kanban board, you can be certain that you are well under way.

While spontaneous quality circles (the Lean term for these ongoing discussions) are excellent vehicles for continuous improvement, many Kanban teams still benefit from the use of a regular cadence of retrospectives (kaizen events in Lean terminology). Retrospectives give the team a chance to gain perspective and see their work from a distance. This sometimes leads to suggestions for larger structural changes beyond Kaizen, which is known as Kaikaku (dramatic change) in Lean. A combination of ongoing quality circles, daily standup meetings and a cadence of retrospectives seem to be a powerful cocktail to drive improvement.

As I mentioned earlier, a key factor in achieving this is to stick to the rule "Change your policies; do not break them". If people do not act according to team policies, chances are that your system will degrade over time and you will never see the true value of visualizing work.

I often get the question "Will Kanban not serve as an excuse to revert to former dysfunctional practices when there are no rules or practices making sure you stick to a more Lean and Agile way of working?" Though I understand where the question is coming from, I do not share this concern. My experience is that Kanban is a unique way of catalyzing Agile and Lean principles, even in situations where it seemed impossible to do so. The few cases where I have seen it fail were either because senior management did nothing to support the initiative or, in some cases, even worked against it or where no effort was put into explaining why visualizing work, flow and feedback are important aspects. To succeed and drive continuous improvement, Kanban needs commitment from management, and people, using the principles in their daily work, need to understand WHY it makes sense. When these aspects are present, there is no reason to expect the initiative to fail or that people will use the system to bypass former change initiatives and revert to old dysfunctional practices.

Good luck on your journey

I hope that these past chapters have given you useful insights and have inspired you to move forward on your Agile journey and to use Kanban on real projects in your company. I would love to get your feedback from reading this and for you to share stories about how it might have helped you achieve better ROI for you and your customers.

All suggestions for a possible second edition are very welcome. You can find me here:

Jesper Boeg

Mail: jbo@trifork.com
Twitter: J_Boeg
Blog: http://triforkagile.blogspot.com/

Trifork Agile Excellence

Mail: triforkagile@trifork.com
Twitter: triforkagile
Website: www.trifork.com

If you think your company could benefit from Kanban training or coaching, we would be more than happy to discuss it with you. Trifork has a broad range of Agile offerings, including coaching, onsite training and certifications in Kanban, Scrum, Lean, Personal effectiveness and Agile development practices.

If you do get started with Kanban, I will encourage you to join the Kanban yahoo groups “kanbandev” and “kanbanops” as well as participate in discussions and knowledge sharing about applying Kanban in practice. Since this only serves as a short intro, I would also suggest that you broaden your knowledge through training and reading the following books:

- Kanban, David J. Anderson, 2010

- The Principles of Product Development Flow: Second Generation Lean Product Development, Donald G. Reinertsen, 2009

- The Elegant Solution: Toyota’s Formula for Mastering Innovation, Matthew May, 2008

- Lean Thinking: Banish Waste and Create Wealth in Your Corporation, James P. Womack and Daniel T. Jones, 2003

- The Toyota Way: 14 Management Principles from the World’s Greatest Manufacturer, Jeffrey Liker, 2004

- The Lean Startup: How Constant Innovation Creates Radically Successful Businesses
  Eric Ries, 2011

Best of luck on your journey and please consider making us part of it.

When I recently received a draft of Jesper's book for review, I didn't think it possible for such a small book to contain both a concise overview and a practical, step-by-step worker's introduction. As I read along, though, I realized that Jesper had succeeded at both. He has a unique gift for getting to the core of a matter, and then helping others get to the core of it for themselves.

I heartily recommend "Priming Kanban" as a practitioner's quick-start introduction to using Kanban. It will enable you to try out Kanban more quickly and painlessly…and defuse your anxiety about doing so.

From the foreword by James Sutton

TRIFORK AGILE EXCELLENCE

Since 1996 Trifork has been pushing the adoption of Agile and Lean principles in IT and has played a key role in the world wide adoption of Agile and Lean in the last decade. Trifork has vast experience in helping large and small companies transition to a more effective software delivery cycle and remains a leading edge provider of conferences, training and coaching in Lean and Agile product development. Trifork offers a wide range of standard and customized courses, workshops and change programs covering everything from low level Agile development practices to large scale Agile and Lean change initiatives. So no matter if you are looking for courses in TDD, Scrum or Kanban, need an introduction for top management to Lean product development or the most experienced onsite coaches guiding your Agile journey - Trifork has an offering that will fit your particular need.

We hope that you will enjoy this book in the Trifork Agile Excellence mini-book series and that you will keep us in mind should you need assistance in your Agile and Lean transition.

Email: triforkagile@trifork.com
Twitter: @TriforkAgile
WebSite: www.trifork.com/trifork-agile-excellence
Blog: triforkagile.blogspot.com
Phone: +45 8732 8787

InfoQ
ENTERPRISE SOFTWARE DEVELOPMENT SERIES

software pilots
TRIFORK.