

MANAGEMENT REVIEW

Management without theatre.

OPERATIONS

KANBAN: THE LIMIT ON WORK IN PROCESS

Toyota's supermarket method of 1954, a BBC Worldwide team with limits on its board, pull as a cap on open work, judges in Rome who juggled fewer cases, four flow measures, and a card for one board.



37%

shorter average lead time in 12 months for a nine-person software team at BBC Worldwide that worked with kanban boards and limits on work in process.

MIDDLETON & JOYCE, 2012

INSIDE

COVER STORY

The next process is the customer

PAGE 03

WHAT THE RESEARCH SAYS

Fewer cases at once

PAGE 06

TOOL OF THE ISSUE

One board, with limits

PAGE 08

NO. 66 · OPERATIONS

In this issue

A board full of cards looks like a busy team. Often it is a team that has started more than it can finish. This issue is about kanban and the limit it puts on work in process: how many things may be open at once, and what happens when the limit is reached.

In 1954 Toyota let each process take parts from the one before, instead of having them pushed on. Half a century later, a team at BBC Worldwide put limits on its boards and delivered sooner and more evenly. Two researchers of operations argue that pull is, in essence, a cap on work in process. In Rome, six judges who kept fewer cases open at once finished them sooner. The Kanban Guide names four measures, and a card helps set up one board.

Stiven Janaqi EDITOR · STIVEN CATALYST

03 COVER STORY
The next process is the customer

04 THE NUMBERS
Limits on the board

05 THE MODEL
Pull is a limit

06 WHAT THE RESEARCH SAYS
Fewer cases at once

07 HOW IT IS MEASURED
Four measures, one oldest card

08 TOOL OF THE ISSUE
One board, with limits

09 SOURCES
Sources and method

HOW TO READ THIS ISSUE

Figure

Every figure has its source and year at the foot of its page.

Our reading

Where the editors interpret rather than the research, it says so.

Practice

The steps and the card are proposals to try, not research results.

COVER STORY

The next process is the customer

According to Toyota's own history, each part in its machining plant once travelled with a voucher and was carried to the next plant whether that plant needed it or not. Parts piled up, and assembly built whatever it had parts for.

Early in the month, assembly reached roughly half the plan. At the end of the month, with all parts in, the conveyors ran faster to catch up, with the same people who had waited for parts at the start. In 1954, as volumes grew, Toyota introduced what it called the supermarket method.

“The subsequent process is the 'customer' and the prior process is the 'supermarket'.”

01**Supermarket**

the earlier process: it holds parts and replaces only what was taken

02**Customer**

the later process: it takes what it needs, when it needs it

03**Kanban**

the card added later: product name, number and quantity

OUR READING

Push sends work on because it is finished. Pull starts work because there is room for it.

Source: Toyota Motor Corporation, 2012

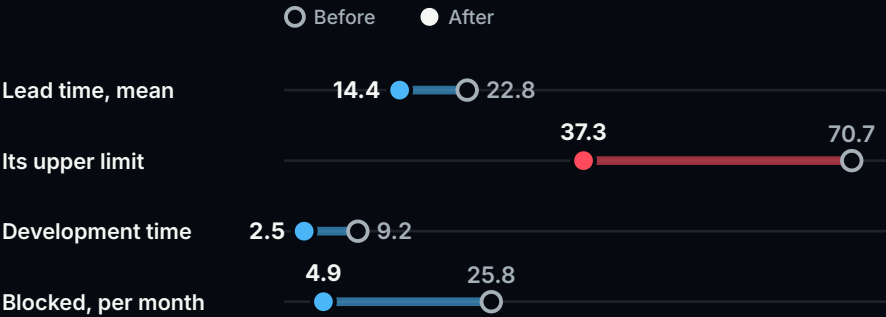
Toyota's history says the idea came from a 1954 trade-press report on a Lockheed aircraft plant, when supermarkets were still rare in Japan. Other tellings credit Taiichi Ohno's look at US supermarkets; we could not check his book, so we follow Toyota.

THE NUMBERS

Limits on the board

In April 2008 a nine-person software team at BBC Worldwide drew its stages on kanban boards and put each piece of work on a card. There was more work in process than anyone had realised. Peter Middleton and David Joyce followed the team for a year.

WORKING DAYS, ONE BBC WORLDWIDE TEAM, 2008–2009



The limits came from the constraints: with fewer testers and analysts than developers, they followed what those could handle, and moved when the constraints moved. Defects reported by customers fell by 24%.

OUR READING

The limit did not make anyone work faster. It made the queue visible, at the stage where it formed.

Source: Peter Middleton & David Joyce, IEEE Transactions on Engineering Management 59(1), 2012 (via Accepted version, Queen's University Belfast research portal)

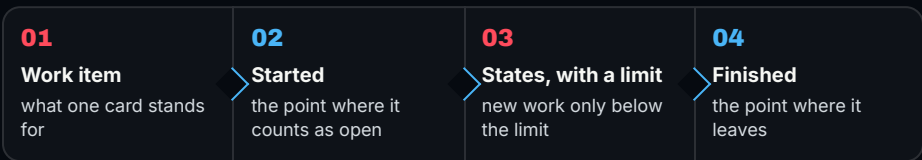
Upper limit: the upper control limit of lead time, a measure of spread (–47%). Development time covers nine months. One case; the team also cut work smaller and grew in skill.

THE MODEL

Pull is a limit

Wallace Hopp and Mark Spearman argued in 2004 that the practitioner literature had simplified “pull” until it was misunderstood. Their definition is short: pull is essentially a mechanism for limiting work in process. A kanban card is one such mechanism: no free card, no new start.

A WORKFLOW WITH A LIMIT, AFTER THE KANBAN GUIDE



David Anderson dates kanban in knowledge work to 2004, when he designed a pull system for a maintenance team at Microsoft. It had no physical board. In his telling, the cards went up on a whiteboard at Corbis in 2007, at a team member's suggestion. His book Kanban followed in 2010.

OUR READING

A board without limits is a picture of overload. The limit is what turns it into a pull system.

Sources: Wallace J. Hopp & Mark L. Spearman, Manufacturing & Service Operations Management 6(2), 2004 (via Abstract at Crossref); John Coleman & Daniel Vacanti, 2025; David J. Anderson, 2014

The definition is Hopp & Spearman's; the chain follows the elements of a workflow in the Kanban Guide (2025). The history is Anderson's own account. The reading is the editors'.

WHAT THE RESEARCH SAYS

Fewer cases at once

Decio Coviello, Andrea Ichino and Nicola Persico studied Italian judges who received the same workload by lottery. Those who juggled more trials at once, instead of working on a few in sequence, took longer to complete their cases. The quality of rulings, measured by appeals, did not suffer.

Then came a test in Rome's labour appeals court. From January 2011, six judges in one section scheduled all the hearings a new case would need up front, so that cases finished before new ones crowded in. 44 judges in four other sections kept the old way. The data run from 2005 to 2014.

12%

shorter average case duration, against the other judges

111 days

less per case, by the authors' estimate

OUR READING

Nobody worked harder. The judges started less at once, and finished what they had started.

Sources: Decio Coviello, Andrea Ichino & Nicola Persico, *Journal of the European Economic Association* 13(5), 2015 (via Abstract at RePEc; NBER Working Paper 16502, 2010); Robert L. Bray, Decio Coviello, Andrea Ichino & Nicola Persico, *Manufacturing & Service Operations Management* 18(4), 2016 (via Abstract at Crossref; working paper of March 2016)

Not a random assignment: the section's president asked to take part. A difference-in-differences study with six treated judges; a placebo test supports it. Their rulings were also appealed less often. Not every study finds lower WIP better on every measure (see No. 58).

HOW IT IS MEASURED

Four measures, one oldest card

The Kanban Guide by John Coleman and Daniel Vacanti asks every kanban system to track four flow measures, each counted between the points where the team says work starts and finishes.

01	Work in process	items started but not finished
02	Throughput	items finished per unit of time
03	Work item age	time since an open item started
04	Cycle time	time from start to finish

The guide pairs them with a service level expectation, such as “85% of work items will be finished in eight days or less”. At BBC Worldwide the daily stand-up, about 15 minutes at the board, checked the cards, then blocked work, then clusters, then priorities.

HYPOTHETICAL EXAMPLE, A RETURNS DESK BOARD ON A MONDAY

Open	14 cards against a limit of 12
Oldest	11 days; 85% usually close within 6

The average looks fine; the oldest card and the overrun show where to act. The numbers are invented.

Sources: John Coleman & Daniel Vacanti, 2025; Peter Middleton & David Joyce, IEEE Transactions on Engineering Management 59(1), 2012 (via Accepted version, Queen's University Belfast research portal)

The four definitions and the example of an expectation are from the Kanban Guide; the stand-up from Middleton & Joyce. How the three averages relate is the subject of No. 58. The example is the editors'.

More in the essay

Ten minutes before the shift: what the team needs to know before the work starts

TOOL OF THE ISSUE

One board, with limits

One team, one flow of work. Count what is open today before you set any limit, start the limits near that count, and lower them only after a review.

01 Work item and flow what one card stands for; where work starts and finishes

02 Columns and limits each stage, today's count, its limit

03 Pull rule who starts new work, and only below the limit

04 Exceptions what may jump the queue, and who decides

05 Blocked items marked on the card; who removes the block, by when

06 Weekly review oldest card, finished items, is each limit still right

Sources: John Coleman & Daniel Vacanti, 2025; Wallace J. Hopp & Mark L. Spearman, Manufacturing & Service Operations Management 6(2), 2004 (via Abstract at Crossref); Peter Middleton & David Joyce, IEEE Transactions on Engineering Management 59(1), 2012 (via Accepted version, Queen's University Belfast research portal)

A practice proposed by the editors, after the Kanban Guide, Hopp & Spearman and the BBC Worldwide team, which set its limits from its constraints and changed them as they moved.

Open the tool: [Shift Handover](#)

stivencatalyst.com/tools/shift-handover/

SOURCES AND METHOD

Every figure has a source.

The figures in this issue come from the sources below. The year shows how recent each one is.

75 Years of Toyota: Supermarket Method

Toyota Motor Corporation, 2012

https://www.toyota-global.com/company/history_of_toyota/75years/text/taking_on_the_automotive_business/chapter2/section7/item4.html

Lean Software Management: BBC Worldwide Case Study

Peter Middleton & David Joyce, IEEE Transactions on Engineering Management 59(1), 2012 · via Accepted version, Queen's University Belfast research portal

<https://doi.org/10.1109/TEM.2010.2081675>

To Pull or Not to Pull: What Is the Question?

Wallace J. Hopp & Mark L. Spearman, Manufacturing & Service Operations Management 6(2), 2004 · via Abstract at Crossref

<https://doi.org/10.1287/msom.1030.0028>

A Brief History of Kanban for Knowledge Work

David J. Anderson, 2014

<https://djaa.com/brief-history-kanban-knowledge-work/>

The Inefficiency of Worker Time Use

Decio Coviello, Andrea Ichino & Nicola Persico, Journal of the European Economic Association 13(5), 2015 · via Abstract at RePEc; NBER Working Paper 16502, 2010

<https://doi.org/10.1111/jeea.12129>

Multitasking, Multiarmed Bandits, and the Italian Judiciary

Robert L. Bray, Decio Coviello, Andrea Ichino & Nicola Persico, Manufacturing & Service Operations Management 18(4), 2016 · via Abstract at Crossref; working paper of March 2016

<https://doi.org/10.1287/msom.2016.0586>

The Kanban Guide

John Coleman & Daniel Vacanti, 2025

<https://kanbanguides.org/english/>

EDITORIAL METHOD

Each figure was checked for its year, its publisher and what exactly it measures. Where the publisher's page could not be opened, the figure was checked against independent summaries and is marked “via”. The editors' interpretation is marked “Our reading”. Figures that could not be confirmed are not in the issue.



MANAGEMENT REVIEW

Management without theatre.

Every issue, one management question, checked against the best research.

ALL ISSUES

stivencatalyst.com/magazine/management-review.html