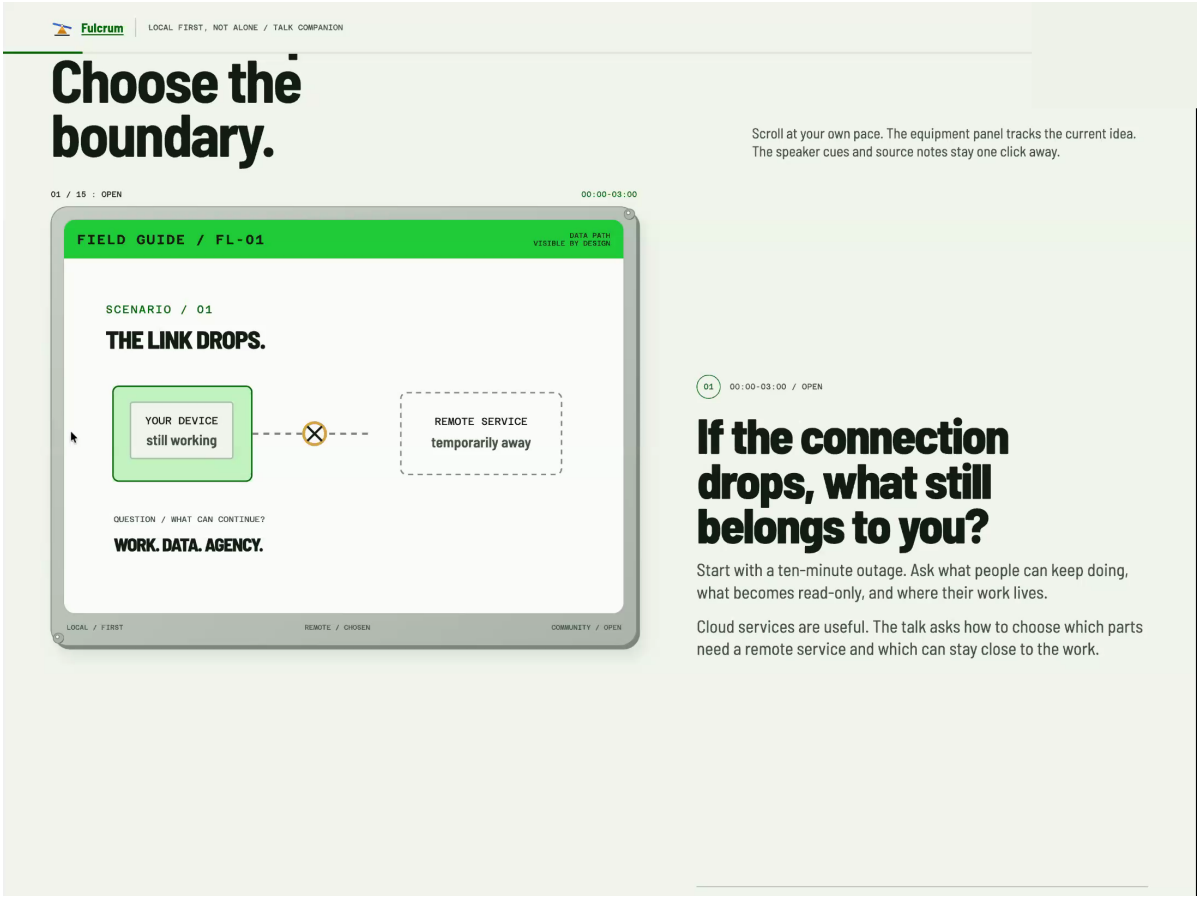
Fulcrum: a local-first agent control plane

Mo Abualruz • 2026-09-26 • English

Full presentation and discussion

Edited transcript of the full presentation and Q&A. Revised on 29 September 2026 against the Soniox transcript, Zoom captions, and presentation material; the Chinese version was checked paragraph by paragraph against the English. Repetitions and broken sentences have been edited for reading, while examples, qualifications, and the speaker's claims are retained. The source recording begins during the talk; the missing opening has not been reconstructed. Timestamps refer to the edited video. Selected unclear audio excerpts were retranscribed with Soniox. Unresolved wording is marked. This is not a word-by-word manual listening audit.

What still works when the connection drops



Recording frame • 00:01:00

00:00:00 • Mo Abualruz

And then the internet cuts out. What do you do? Stop working? Stop making progress? This is where local-first system design comes to mind: I need at least read-only access to my data so I can continue working during an outage. I wouldn't say there is one answer, such as “Build everything in the cloud” or “Build everything locally.” You have to cover the edge cases so that no point of failure is catastrophic. You should be able to recover from failures and keep working through blockages or downtime. That is an important way to think about your own setup.

00:01:07 • Mo Abualruz

This also applies to software design: you can separate things into layers and tiers. But it applies to your work, too. You want an environment that keeps you from getting blocked and lets you continue as much as possible. If you have other approaches or questions about local-first and self-hosting, we can discuss them.

00:01:47 • Mo Abualruz

I think of it this way: I can host things that complement the cloud and provide failover during outages.

00:02:00 • Mo Abualruz

The design idea is this: Keep a copy of whatever you critically need from the internet or cloud providers. You can also treat your home lab as a cloud provider. For example, you don't necessarily need Jira if you have a good deployment, with backups, of a task manager such as Plane. Plane is a good open-source project.

00:02:33 • Mo Abualruz

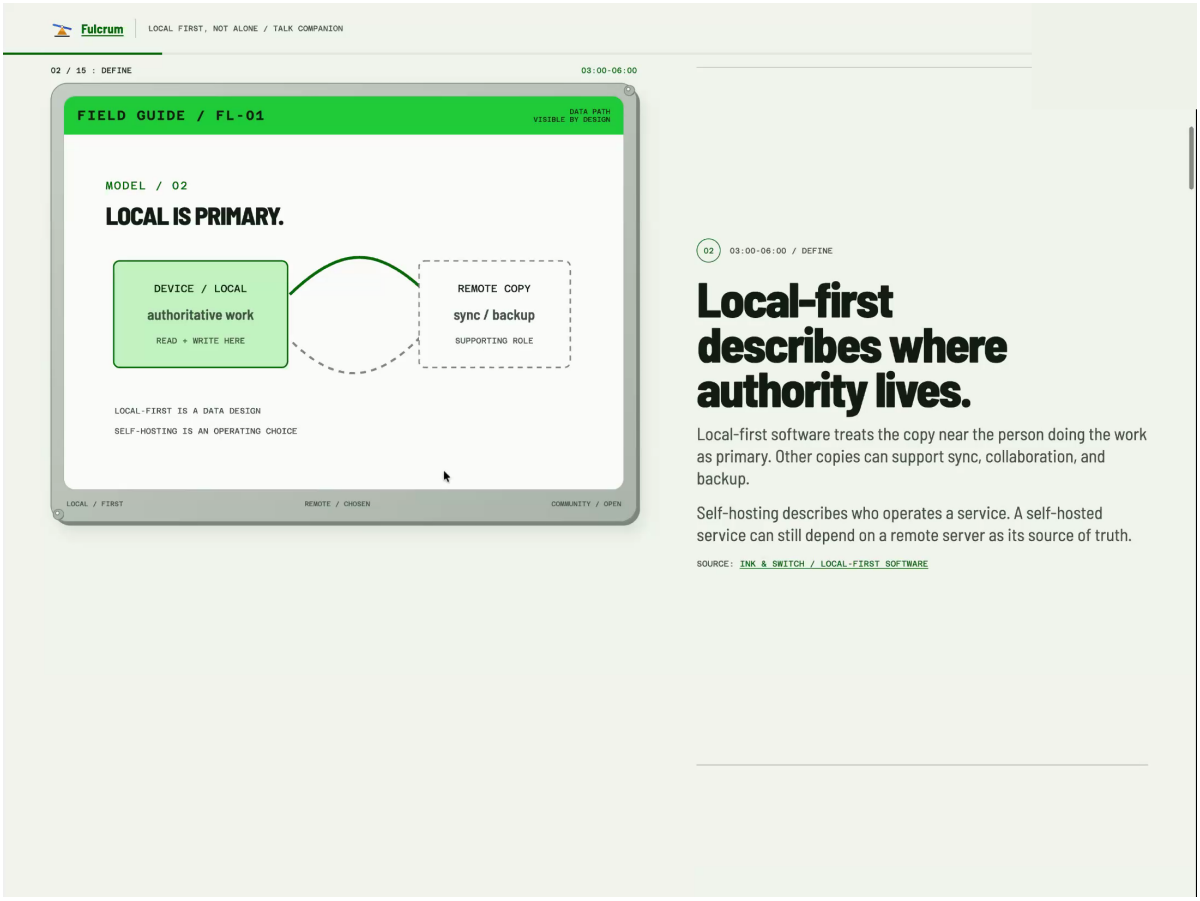
It has almost every feature Jira has, but it is open source. You can self-host it, integrate with it, and add friends or colleagues to work on projects and planning. But that is not the whole picture.

00:02:51 • Mo Abualruz

If something happens to the machine running your service, you also need things such as Markdown or JSON files to track your work. You can still read them and see the status: “Before the service went down, this is what we were working on.” That is the local-first mentality: keep copies that let you stay unblocked. I self-host most services, but I also use tiers. A local LLM does certain work and manages my memories. If the local service fails, I can consider using a cloud model. [The spoken sentence about cloud connectivity is contradictory; the fallback intention has been retained here.]

00:03:35 • Mo Abualruz

For each person, the important thing is to identify where you most often get blocked and design a way to keep working. Keep your momentum, stay engaged with the project, and understand what is happening. In CLI or chat sessions, you often get a textual description: “We are doing steps one, two, and three.” I add graphs to clarify that, because otherwise it can become hard to follow the session fully. Most of the work that keeps you unblocked is read-only work, so that information needs to be available on your device. Your home lab is also a provider: if you leave home and need to work, you should have local data so you can continue without relying even on the home lab.



Recording frame • 00:04:30

00:05:00 • Mo Abualruz

Sometimes you are traveling and still want to get things done. There is a good paper about local-first design. Let me share it in the chat. Is it full screen? This is a good paper to share. I can also share the presentation after the session; it contains all the references I am using.

00:05:38 • Mo Abualruz

The paper explains local-first software and how to think about it. Essentially, it is about having your own copy. These days we carry credit cards, but it is still useful to carry some cash. Not every shop, city, or place is connected.

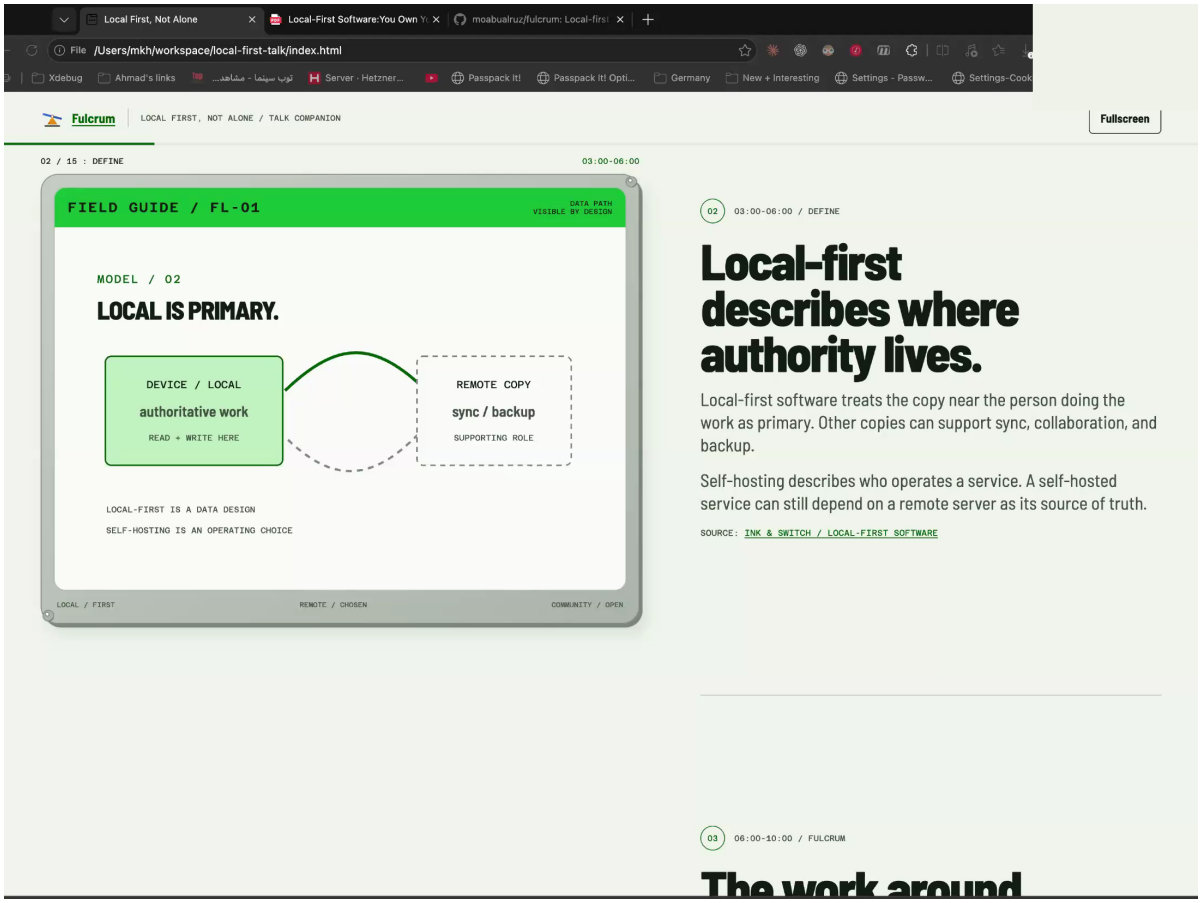
00:06:05 • Mo Abualruz

Not every country has the same infrastructure either. Think of each cloud or self-hosted service as a credit card, but keep some cash just in case. If something goes wrong, you still have what you need. When nothing goes wrong, the same system should give you the same data, so you don't have to interact with multiple systems to obtain it. You have one communication path that manages your work.

00:06:48 • Mo Abualruz

That is why I unify communication around the authoritative information: the important data I need to read, the context, memories, and code. What matters is not simply what is on GitHub. The work on your machine may be the most recent version. Your local copy has the new tasks, the features you are working on, and your ideas.

The work around the agent



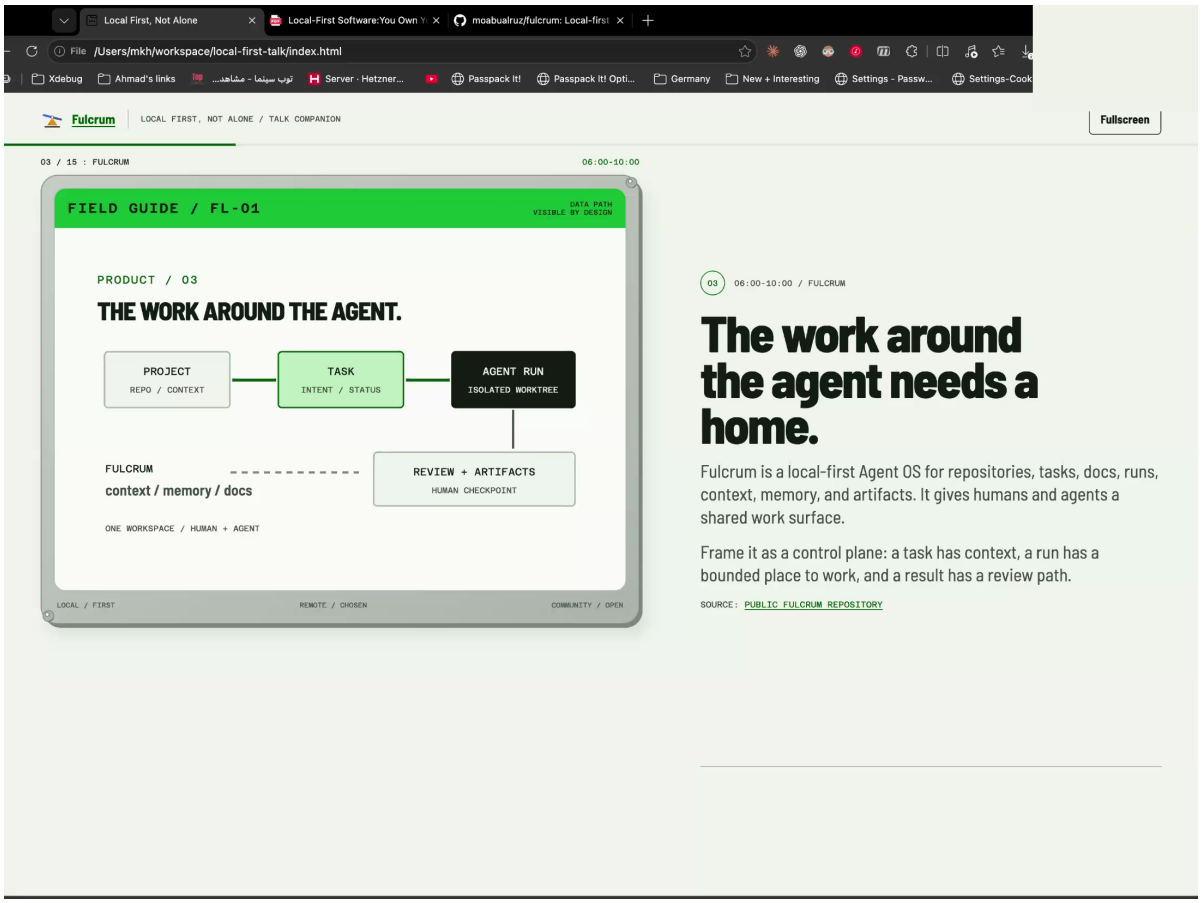
Recording frame • 00:07:30

00:07:23 • Mo Abualruz

Everything is documented there. If GitHub goes down, your machine still has the code and files. But suppose you rely on Honcho, an open-source memory-management project, through its hosted service. If you lose internet access, you cannot query your memories. You might also want memory files, or a local Honcho instance with a small model that copies what you have in the cloud.

00:07:58 • Mo Abualruz

That local model might not be as powerful as the hosted one, but at least you have a copy and can continue working. This is the idea behind local-first and how Fulcrum was built. For agentic workflows, I break work down by project and task, and then each agent run is a unit of work.



Recording frame · 00:10:30

00:08:31 · Mo Abualruz

A task can be planning or review. Coming back to Fulcrum, at some point you reach the review stage. I review my own work because I am technical. If you are less technical and ask Qwen to do a task, you can ask Kimi [version unclear] to review it, using a more capable model in the cloud. Qwen 27B has already done most of the coding, which makes the process cheaper: use a small model for the task, then a more expensive one for review.

00:09:25 · Mo Abualruz

You might not have the technical experience to find every mistake, or you might be learning a language and benefit from Kimi's feedback: “This is how we do things in Go; that is a bad pattern.” The smaller model makes a mistake, and you learn from the frontier model's review. [The model name is repeated inconsistently in the recording; this sentence follows the surrounding review example.]

00:09:51 · Mo Abualruz

Fulcrum has to provide all the links: memory, documents, and context for the task and review. It is not enough to say, “Here is a PR; review it.” You cannot just give that to an agent or a person without context. In my experience as an engineering manager, we always ask, “What are you working on? What is the value of this PR? What problem are you trying to solve?” That information matters to a human and to an agent doing a task or review. They need to understand what you want to achieve. Random code changes or token generation are not really work; you need to make sure the work produces something of value.

00:10:44 • Mo Abualruz

Explaining that makes things easier. In Fulcrum, I guide agents through a kind of test-driven design, but it is more value-driven. I don't emphasize unit tests as much as value tests, which are end-to-end tests.

00:11:10 • Mo Abualruz

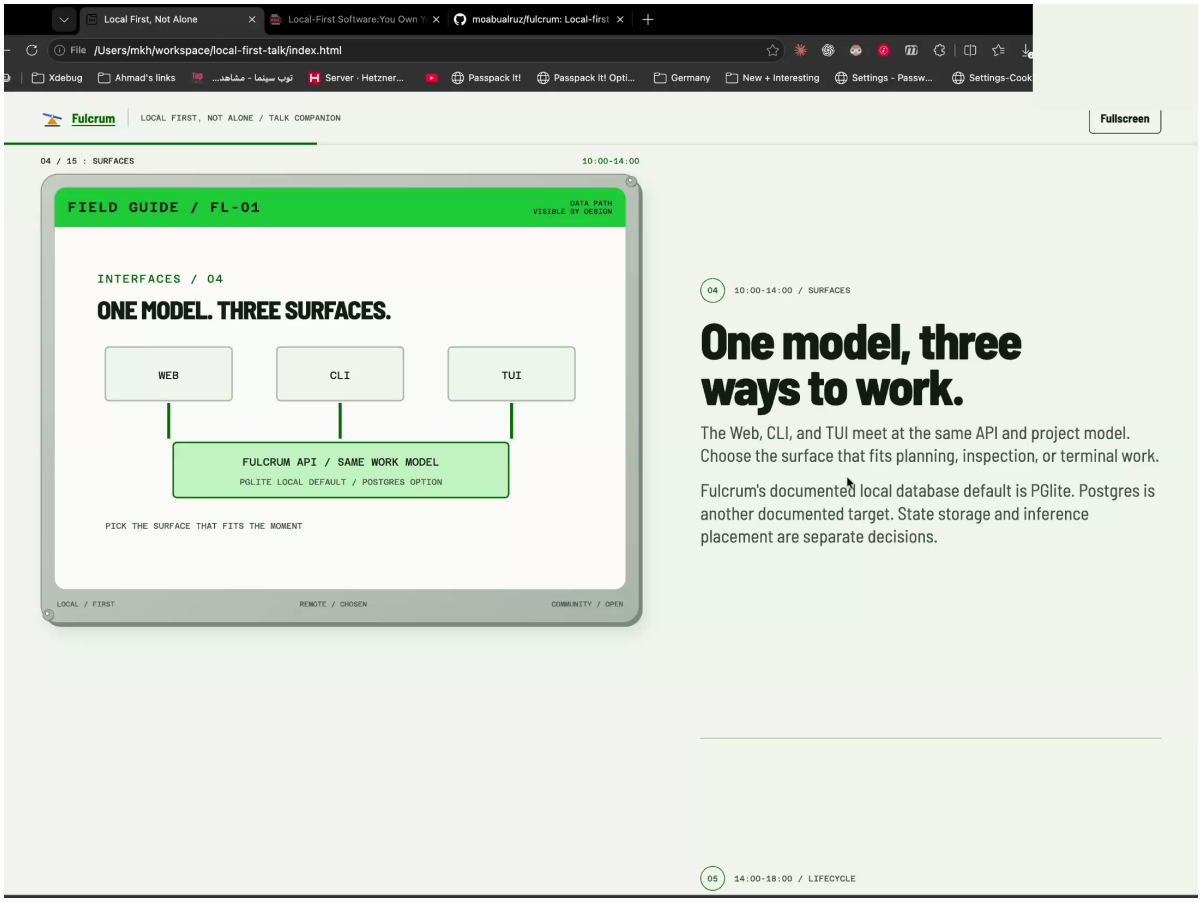
I want the agent to check that the user workflow and user story are correct, and that the value I described is testable in a workflow. My expertise is in code, so I want AI to complement that expertise. I can review code and guide the technical work. The most valuable contribution from AI for me is testing: can it check the user's edge cases and all the changes? Sorry, I have a phone call. I try to add tools for more use cases in Fulcrum, covering more of the agentic workflows people need.

00:12:11 • Mo Abualruz

It is an open-source project. The features I added complement my skills, but the host might have different skills and need additional features for his workflow. He could add one or two things that complement the way he works. That is why I think projects that provide context and observability for your work should be open source, rather than remain only part of one person's journey or a private repository.

00:13:02 • Mo Abualruz

That is also why Fulcrum has three ways to interact with it. The TUI is for people who prefer a terminal and want visibility there. The CLI is for a person or an agent that wants to perform actions. The web interface lets us see and interact with the full system.



Recording frame • 00:13:50

00:13:31 • Mo Abualruz

They use the same system, API, and calls, with three presentation layers. These can serve people, agents, and coworkers, including multiple people assigned to different activities involving the agents.

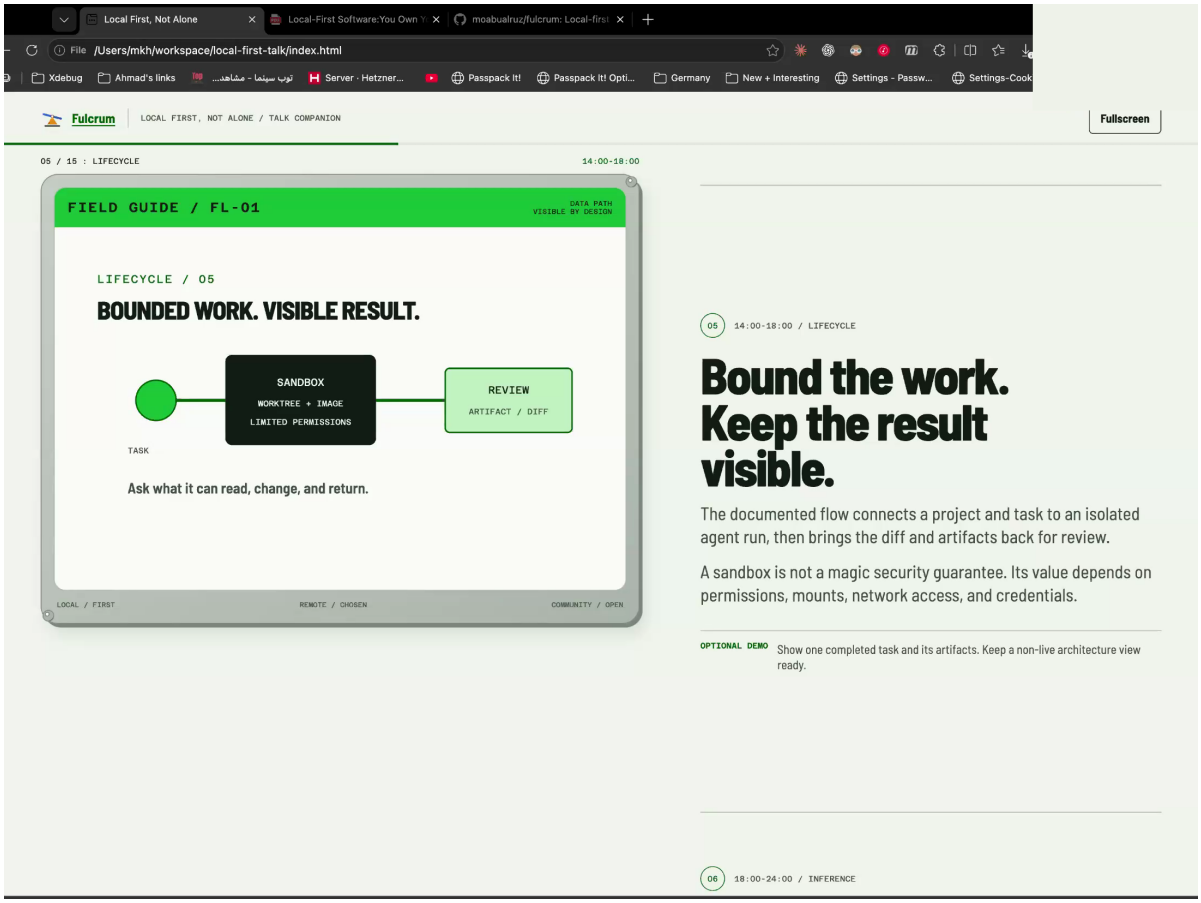
00:14:03 • Mo Abualruz

Communication is JSON across the board. Without going too deep into the technical details, the interfaces use gRPC for faster API calls from the CLI and TUI. It feels event-driven: data can be pushed to the frontend and sent back to the backend, giving you two-way communication. It is faster than the usual approaches. I don't consider it as reliable as full HTTP for handshakes, and I wouldn't use it to integrate with something remote.

00:14:48 • Mo Abualruz

But when something runs locally with your own interface, the benefits of speed, clarity, and encapsulated processes are great. That is how I could build three interfaces for Fulcrum around one system.

Bounded work and visible results



Recording frame • 00:16:00

00:15:08 • Mo Abualruz

How we work in Fulcrum matters a lot to me. I don't particularly like hard sandboxes, but I want more work to happen in parallel. I want agents and people to have freedom to edit without breaking existing work. That is why the sandboxing is built around Git worktrees.

00:15:36 • Mo Abualruz

An agent or I can create a new Git worktree and experiment or prototype freely without damaging the main checkout. After review and approval, we can merge that work into the main local branch.

00:16:02 • Mo Abualruz

This is all local work. I am not talking about a PR review; I mean reviewing the agent's or your own result before it goes back to the main checkout, before creating a PR. This is a lightweight way to give us freedom while protecting current work and avoiding unintended actions on the machine. I don't want an agent to confuse a command for the work with a command affecting the machine.

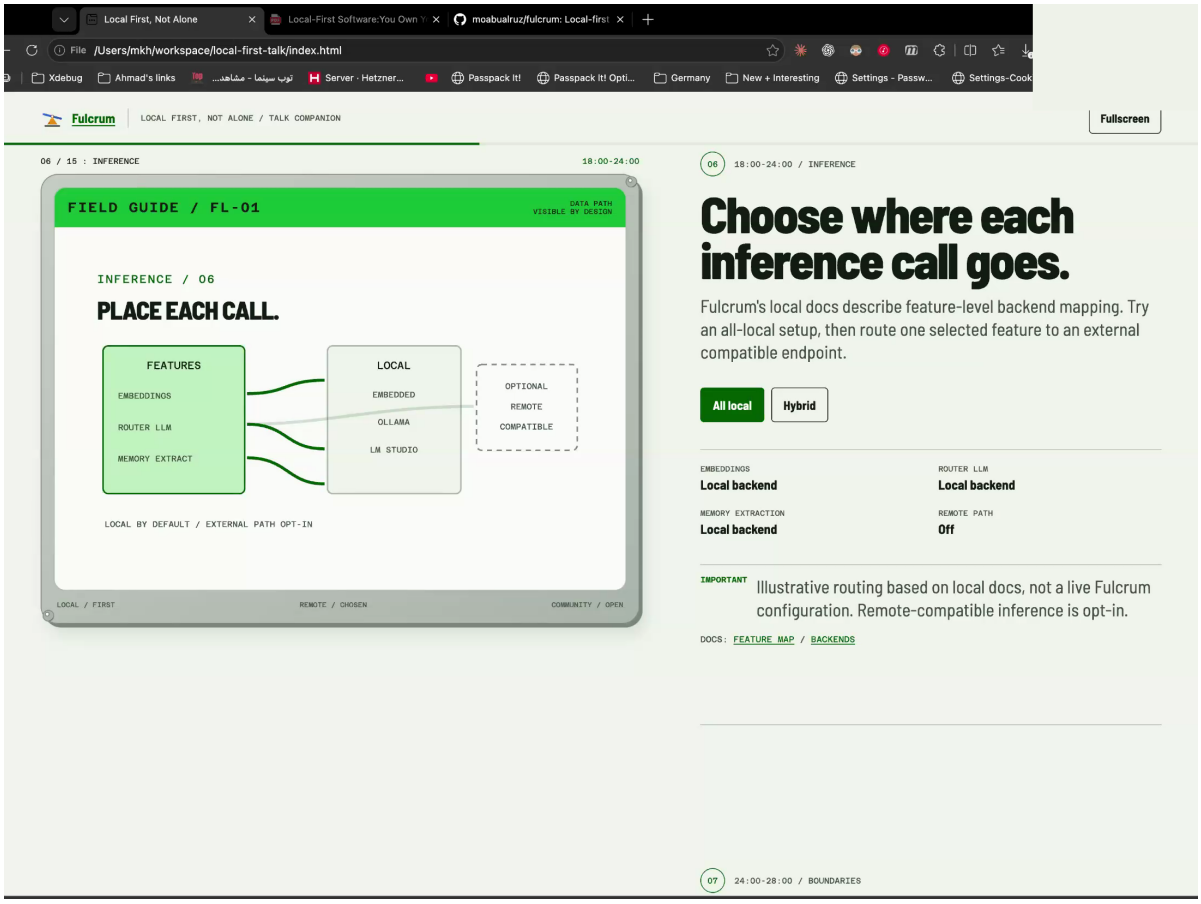
00:16:36 • Mo Abualruz

The Git worktree is the working directory. We can review its output and then put it on the workset branch.

00:16:50 • Mo Abualruz

We went through the integration part quickly. Where was that image? I have lost it, but I can explain. When configuring Fulcrum, you can route every LLM interaction to local models. You can also integrate harnesses to drive the work, with each harness configured as a harness.

Inference routing and traceability



Recording frame • 00:19:30

00:17:31 • Mo Abualruz

A harness can use cloud-only inference or local models; that is a harness-level configuration. Fulcrum does not dictate how you configure it. This gives users flexibility: perhaps they want Kimi for coding but Qwen for planning. Requiring a harness to use the same LLM providers as Fulcrum would be too restrictive. Separately, Fulcrum links documents and tasks, so some of its main features use embeddings.

00:18:39 • Mo Abualruz

I would not call this RAG; it is more like a graph. We build links between context nodes: this file relates to this task and this plan. You can see how the data is connected and how it relates to the work. When we give a task to a harness, that helps us provide guidance, especially for weaker models, which need more of it.

00:19:12 • Mo Abualruz

We can pull the graph links and provide them to the harness when it starts. That is where embeddings come in. You can run embeddings on OpenAI, or use Claude models if you want. But embeddings are very cheap to run locally; most of the models are around three or four gigabytes.

00:19:39 • Mo Abualruz

Because this is not RAG, you don't even have to rerank them. You just need them embedded, linked in a graph, and quick to retrieve. The router LLM essentially chooses where requests go.

00:19:54 • Mo Abualruz

I set routing rules. When planning locally, I try to use Qwen 27B if it is available. For quick research, I don't mind using Gemma 4B. All it has to do is use my web MCP and [search tool name unclear], retrieve search results, and give me a research summary.

00:20:23 • Mo Abualruz

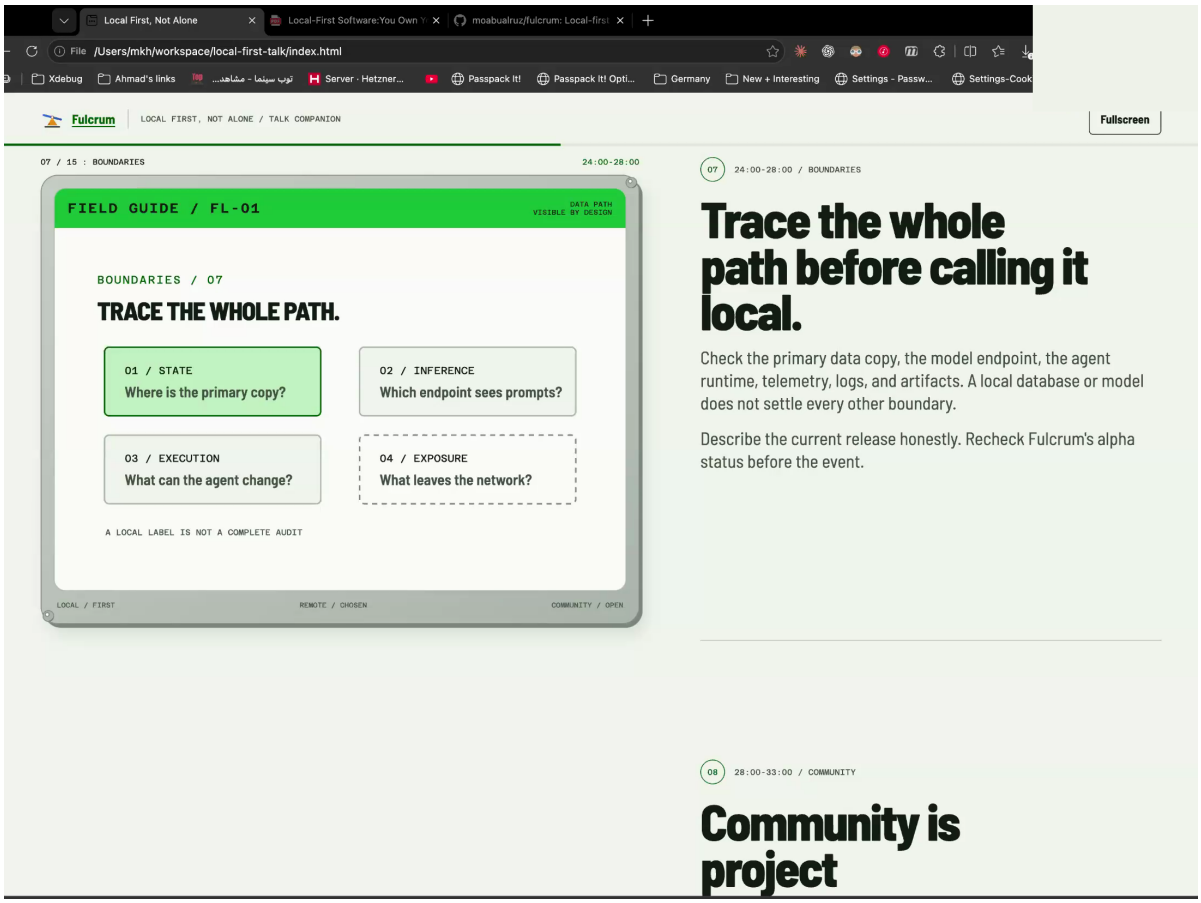
It does not have to be especially smart; it just needs to find what I query on the internet and summarize it. I set those rules in a configuration file. Everyone can configure tier-one, tier-two, and tier-three models and how to run them. They can all be local, or a mixture of local models and cloud inference providers.

00:20:51 • Mo Abualruz

That is how you direct each request or inference. The LLM router can use LM Studio, Ollama, Claude, Kimi, or another OpenAI-compatible endpoint. I have a separate tool for Claude, because its API is not OpenAI-compatible. The tool uses the Claude SDK, connects to your account, and converts the endpoints to an OpenAI-compatible interface.

00:21:39 • Mo Abualruz

That is for my convenience, but native Anthropic support could also be added to Fulcrum, handling Claude's message format and API structure directly. I think MiniMax in China uses the same approach. They interact differently from the OpenAI inference convention.



Recording frame • 00:23:00

00:22:13 • Mo Abualruz

Now, the next part: traceability. This is important for anyone working on a project, not only an AI project. You need traceability and observability to know what is happening and how the work is progressing.

00:22:39 • Mo Abualruz

In some roles, you need to understand machine performance: is it CPU-bound, RAM-bound, or I/O-bound? In Fulcrum, we need to see inference usage. How much inference am I doing? Is it too much? Can I optimize the configuration?

00:23:05 • Mo Abualruz

I might improve execution too. I work on multiple projects with agents locally, and I now make them pull the whole dependency tree into one PR. I found that execution was producing too many requests and too many PRs involving only a few files.

00:23:35 • Mo Abualruz

A review request with one or two files may not capture the context or the value. The AI might see the bigger picture because we give it the graph, but that can be hard for me as a human. Since the graph also links issues and dependencies, I can steer the agents to pull those together.

00:24:02 • Mo Abualruz

Can we put all the dependent issues into one review so it is more meaningful? Without paths and views that let us visualize and interact with the work, it is hard to identify weaknesses we can improve. Traceable paths and views let me see everything and optimize my workflow.

00:24:36 • Mo Abualruz

This is configurable and can be guided by instructions. It is not hard-coded; configuration and Markdown files guide the agents. Rigidly coding the steps might suit my workflow but prevent others from adapting them to theirs. Before I change gears, you can ask questions specifically about Fulcrum. Next I will discuss community: why knowledge-sharing sessions like this matter, and why users and contributors are important to Fulcrum and other open-source projects.

00:25:43 • Mo Abualruz

Before I move on, feel free to ask anything specifically about Fulcrum.

00:25:57 • Host

Thank you so much, Mo. If anyone has questions, you can ask them now.

Community, agency and shared knowledge

00:26:10 • Host

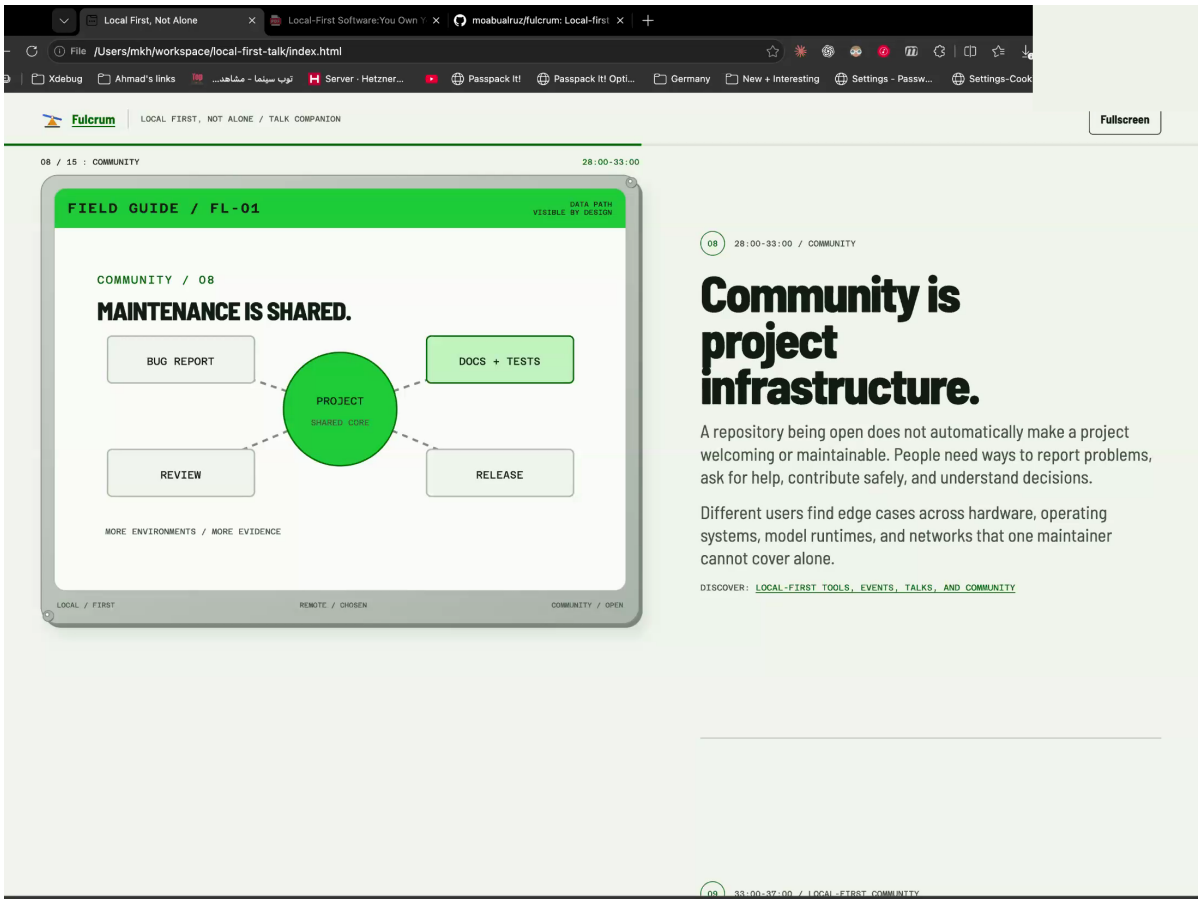
[Translated from Chinese] You can ask questions now. He has explained how Fulcrum works, and next he will talk about why community matters and how it connects with the project. We can first take questions about the project itself.

00:26:45 • Host

[Translated from Chinese] Someone has posted the GitHub link in the chat.

00:26:53 • Mo Abualruz

That's fine; I am sharing the link so you can explore and ask questions. There will be another Q&A at the end too. An important point is that being open source does not automatically make Fulcrum completely welcoming to a community. People need ways to interact with the project. I have tried to make it welcoming, but there may be things to improve: ways to file issues and communicate with me or with future maintainers. This applies to any open-source project.



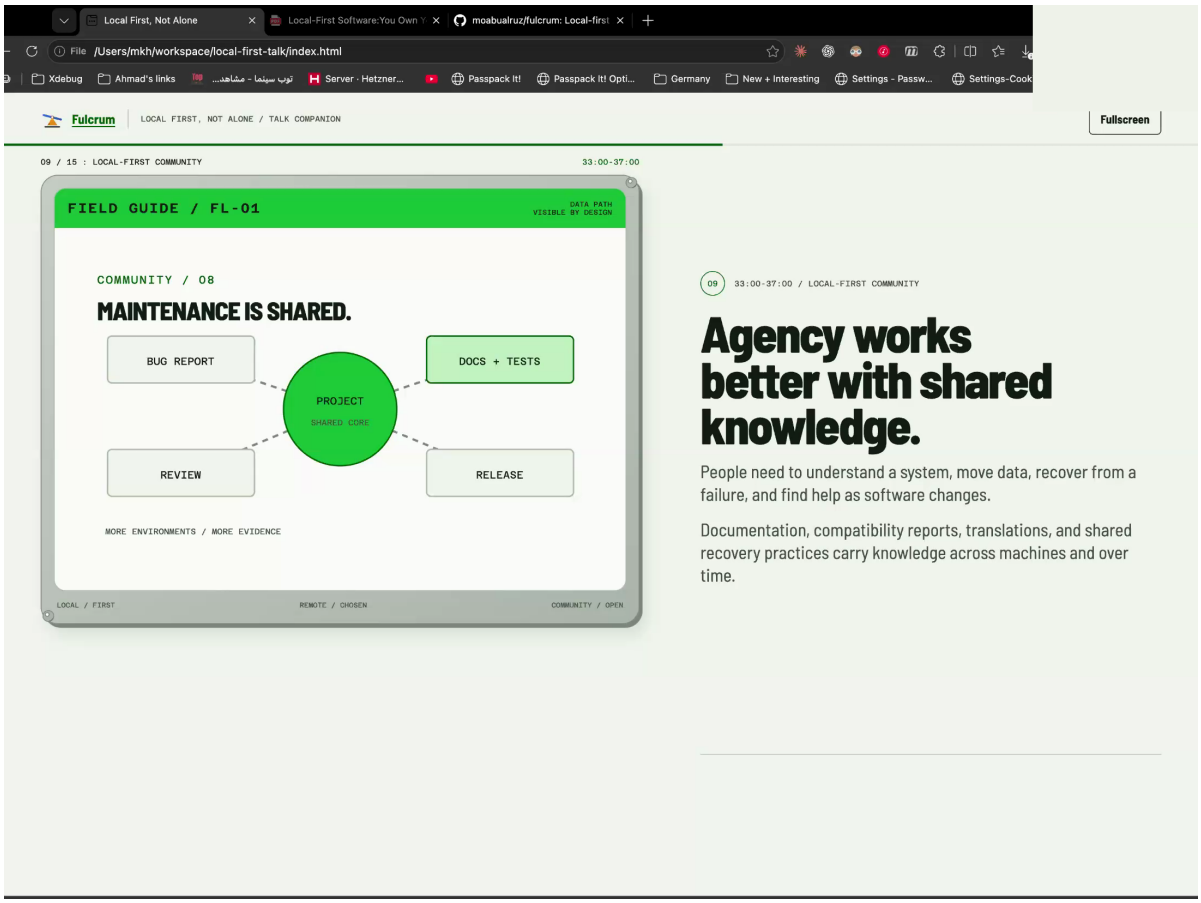
Recording frame • 00:28:20

00:27:48 • Mo Abualruz

Think about this for your own projects too. If I open-source something, I need to build the community as well as the product. Community is another path of growth to maintain. A larger community brings more feedback, which helps refine the product and its ideas.

00:28:19 • Mo Abualruz

People contribute not only through code and issue reports, but through communication and ideas. Events like this let me explain my thinking and hear other people's questions and experiences from their projects. Everyone can share knowledge and benefit. If Fulcrum is to succeed beyond being my own local agentic OS, I need to build a community around it. If that fails, it will still be a useful tool for me, but it will not gain the same adoption, improve at the same rate, or develop the same capabilities. Human agency works better with shared knowledge. Teaching people gives them more choices. An informed decision is better than a decision made without that information; knowledge helps people decide more accurately.



Recording frame • 00:30:20

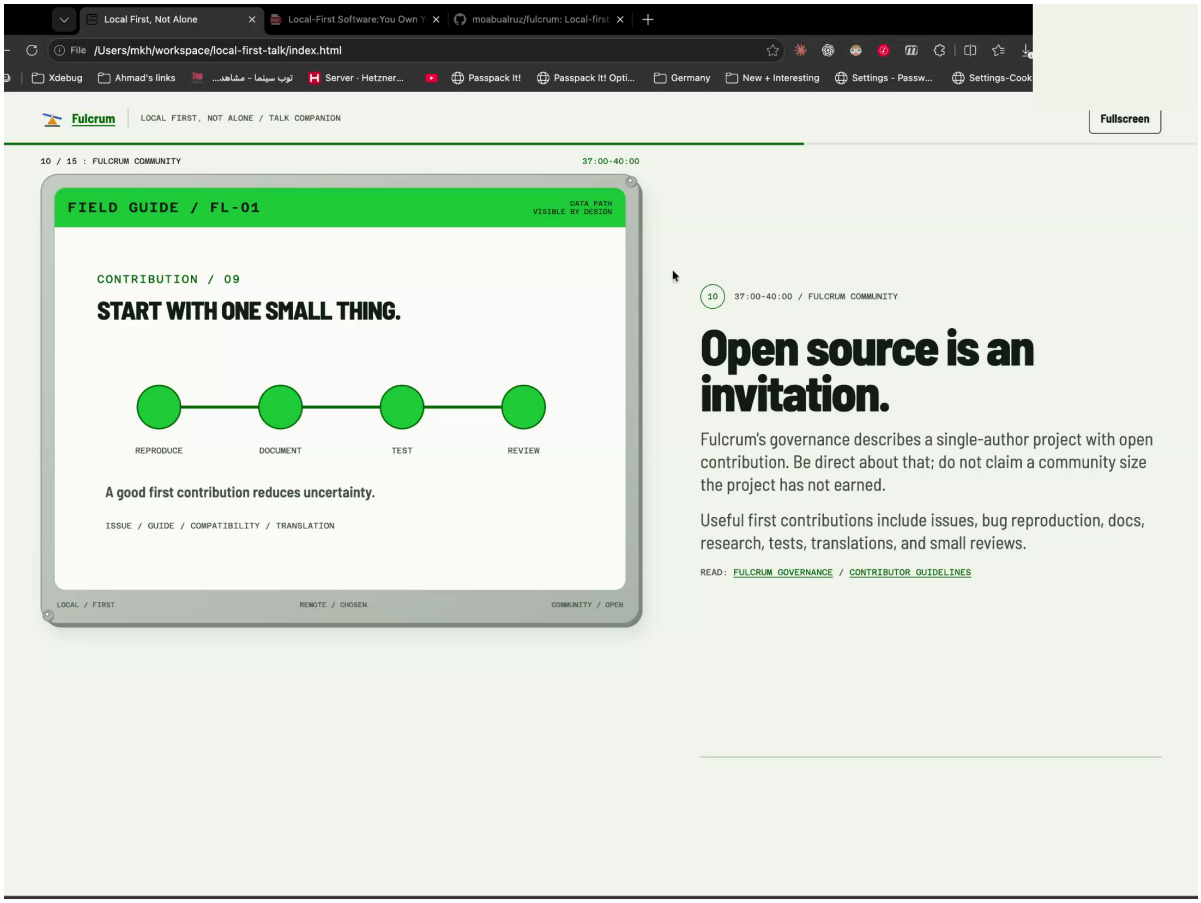
00:29:49 • Mo Abualruz

When I was young, people would sometimes try to trick me: “Would you like these two coins instead of this one?” As a child, two seemed better than one. But I had to learn to turn the coins over and check their value. If nobody taught me to look at what was written on them, I would not know that the two coins were worth one each and the single coin was worth five. I should have chosen the one coin. That is how we learn to make informed decisions. Without agency, freedom to learn, and communication, people can still make decisions, but they will not be as well informed.

00:30:45 • Mo Abualruz

The system should help you stay informed, make good decisions, and move forward. Now, on to contributions.

Documentation, contribution and local services



Recording frame • 00:33:00

00:31:10 • Mo Abualruz

The first two kinds of contribution are technical: reproducing issues and writing documentation. Documentation is where, I would say, many engineers fall short.

00:31:30 • Mo Abualruz

Documentation is key. It is hard and tedious, and not the most fun or engaging part of building something. But it matters, especially in the AI era. It helps people understand the product and how to improve it, and it helps agents retrieve context quickly. Code can be documentation, but not everyone can easily understand it.

00:32:15 • Mo Abualruz

Not everyone has the same technical experience. An agent might not care whether details are in code comments or a documentation portal. But for a new person, perhaps just out of university and still learning, diving straight into code on the first day is hard.

00:32:43 • Mo Abualruz

Documentation guides people and is a foundation for community. I hope Fulcrum will include user manuals and other kinds of documentation too. Another part is testing. How can contributors or users validate our claims? How do we know that everything really works, and that new additions work with what was already there? We need tests.

00:33:25 • Mo Abualruz

As the community grows, you and the community also need to review work. Reviews build knowledge. Agents can teach you things about a language, as I mentioned, but people can too, and learning from a community can be more engaging.

00:33:52 • Mo Abualruz

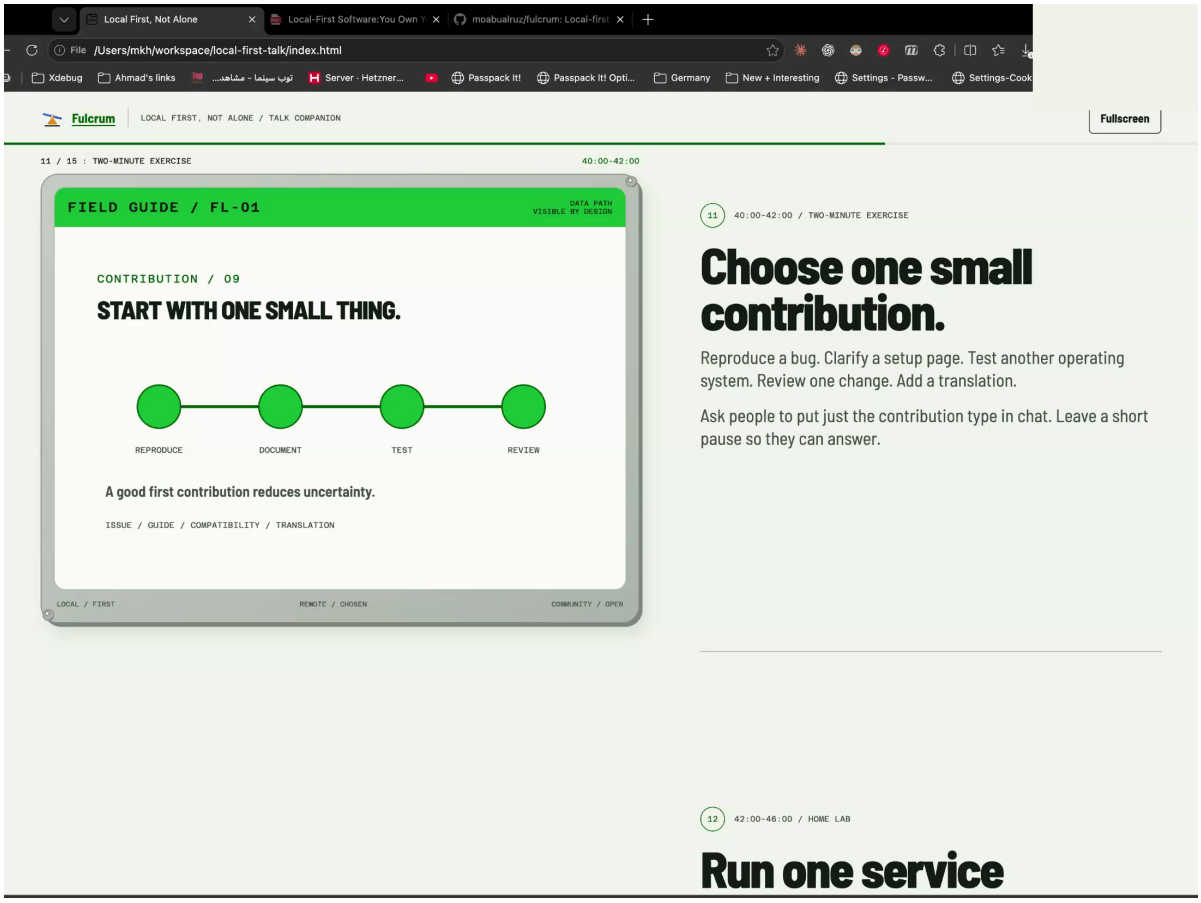
An agent's review may give you a couple of lines saying to change something because it is a bad pattern. In a conversation with a person, you can ask why and get a deeper explanation. Of course, it depends on the person; not everyone's communication skills are equally strong.

00:34:20 • Mo Abualruz

But most people I have met will open up and explain more if you ask why and show that you want to learn.

00:34:36 • Mo Abualruz

I also want to encourage contributions. Just because something seems minor does not mean you should not contribute. Open an issue. This is not only about Fulcrum: in any open-source project, you may find something that wastes time or energy.



Recording frame • 00:35:50

00:35:13 • Mo Abualruz

Perhaps it does not work well, but you have a small workaround and carry on. We do that all the time without sharing what is wrong. Products can then grow in the wrong direction because people do not report their problems. They might even have a local fix but never submit the code or suggest it as a solution.

00:35:42 • Mo Abualruz

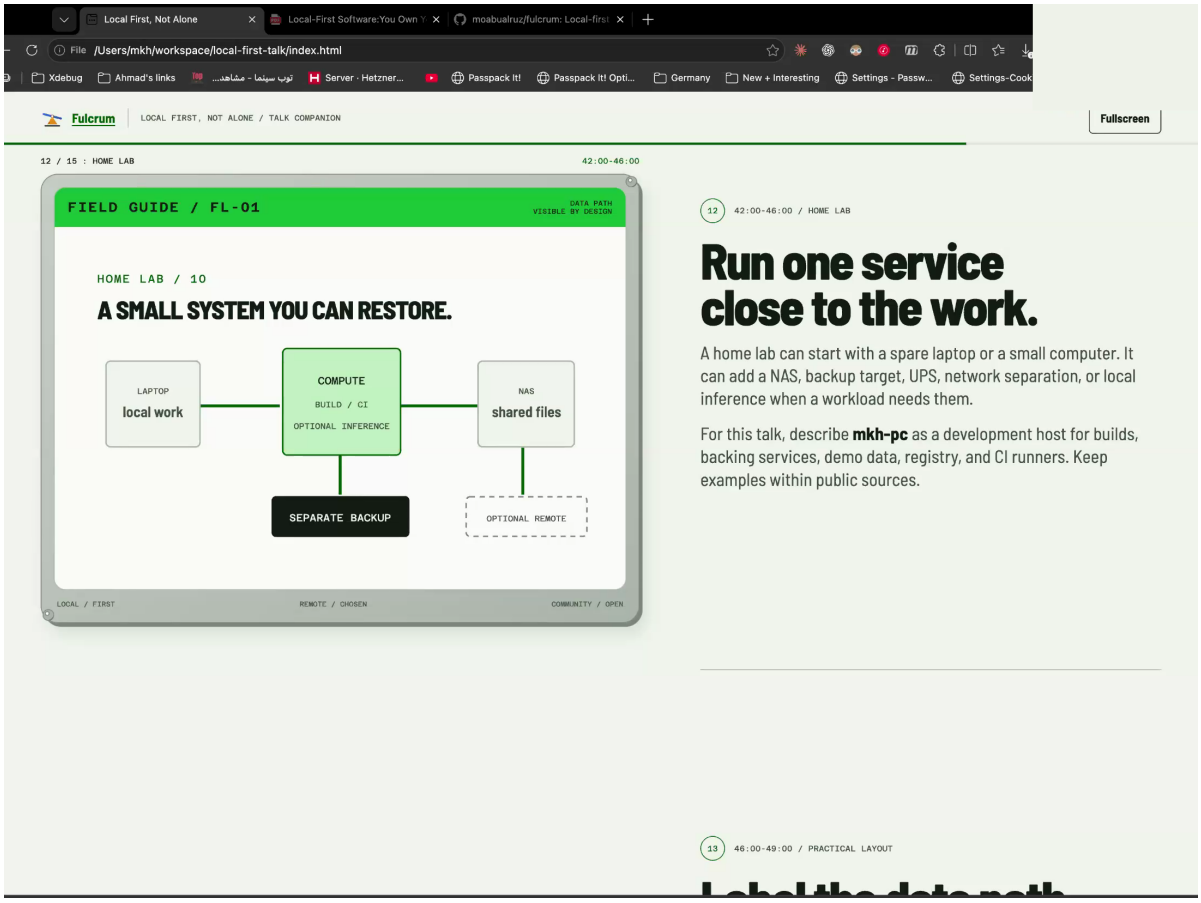
Once you understand a problem and can reproduce and explain it, contribute or report it: “This is happening; this is what I need.” That helps the community grow, and your contribution will be appreciated.

00:36:07 • Mo Abualruz

At least you will learn whether the behavior is intentional. Something that annoys you might exist to address a security issue. By asking, you discover why the decision was made. Pressing an extra button may be inconvenient, but it protects your data.

00:36:34 • Mo Abualruz

Then you can understand and accept the inconvenience. If you never ask, it just keeps bothering you. That finishes the open-source and community part. Next I will talk a little about my own setup and the services around Fulcrum and my development work. Before that, feel free to ask about contribution or community-building.



Recording frame • 00:39:30

00:37:23 • Host

[Translated from Chinese] If you have questions about contributing or the community, you can ask now. Otherwise, he will move on to his current setup.

00:37:48 • Host

Any questions? [Translated from Chinese] You can also look at his GitHub.

00:38:01 • Mo Abualruz

Okay.

00:38:02 • Host

Okay.

00:38:03 • Mo Abualruz

If you remember a question later, there will be time at the end. This next part should take about five minutes. An important point is that no single service, including Fulcrum, can do everything you need. Think about the whole development environment. I have a PC at home running many services. Some disks serve as a general NAS for my wife's and children's photos and other files. I also use it for action runners, because I don't want to pay GitHub $1,000 a month to run actions. They are a series of steps to execute, and I already have a machine. I can direct the work to it. It is easy: create a GitHub App, give it an IP or API, and queue the runner requests.

00:39:17 • Mo Abualruz

Then you can run them on whichever machine you want. It is not a difficult project; it might take a day. Open-source projects already provide ways to manage it. If you lack a public IP, I think you can use NetBird or another personal mesh VPN product so GitHub can access the local machine and queue actions.

00:39:43 • Mo Abualruz

If the machine holds important data, think about a UPS. How much battery capacity would keep it running for an hour during a power outage? I think China's power supply is fairly stable, but here in the Middle East we have frequent outages.

00:40:11 • Mo Abualruz

Local-first means I keep all the data on my laptop, which can run on battery for a long time. But it cannot run all my other services or GitHub Actions; at some point I still need the PC. That is where a UPS helps keep those critical parts running.

00:40:43 • Mo Abualruz

You need to design the network, upload and download capacity, and storage, weighing value against cost. Why did I buy expensive disks for family photos? We had reached the point of paying Apple and Google around $200 a month for extra storage because my wife and children take lots of pictures.

00:41:19 • Mo Abualruz

The alternative was to go through old pictures and start deleting things. But the value of those pictures and memories to my wife and children is... [The speaker pauses to take a question.]

Q&A: models, context and collaboration

00:41:37 • Audience (chat question)

Hi, thanks! Sorry, I may have joined late, but how is this different from an agent running a local LLM, such as a DeepSeek harness running Qwen? I need to persuade our board to switch. Thanks!

00:41:37 • Mo Abualruz

Sorry, there is a question.

00:41:52 • Mo Abualruz

I think this is a question about Fulcrum. You can use any harness with it, and it can also run inference directly through Ollama or another provider. But Fulcrum itself is not a harness. It is a local-first integration layer that keeps your data, connections, and context together.

00:42:19 • Mo Abualruz

You can store memories in Fulcrum files or integrate Honcho or Mem0 as the memory backend. The point is to connect the context across whichever harnesses and workflows you use, so you can move forward and observe the work.

00:42:44 • Mo Abualruz

You can see the steps without switching between Fulcrum and two other products to update plans or tasks. It gives you a platform where you can see, plan, and execute most of your work.

00:43:11 • Mo Abualruz

Have I integrated the DeepSeek harness directly? It is fairly new. I have tested it by itself, but not integrated or tested it with Fulcrum. I made a plugin for Pi, and the DeepSeek harness has similarities to Pi. I think a small plugin could integrate it easily.

00:43:37 • Mo Abualruz

[Reading the next question about tracing and auditing.]

00:43:55 • Audience (chat question)

Does Fulcrum offer trace and audit functions for memory, context, and skills as LLM versions change and the project progresses?

00:43:55 • Mo Abualruz

If you have a skill that runs audits, yes. But if you mean a dedicated audit step, I currently treat auditing as part of review. Sorry, did you say something?

00:44:19 • Mo Abualruz

In my workflow, auditing is part of review. Without the community, I might not see the need for a specialized audit step in Fulcrum. If you need one, open a ticket or issue. If you have the technical skills, you could submit a PR: “I added this feature, tried it, and it works. An audit step fits here.” There is no dedicated audit step at present. I do have a doctor script that can check skills, but those scripts mainly check Fulcrum's installation and harness integrations, rather than auditing memories or context.

00:45:14 • Mo Abualruz

That could be a useful feature. Open an issue for me to look at, or implement it yourself if you have an idea.

00:45:28 • Mo Abualruz

I welcome both.

00:45:32 • Audience (chat question)

Running many agents, sharing context, and maintaining that context sounds tricky.

00:45:32 • Mo Abualruz

It can run multiple agents. You configure how many can work concurrently. By default, each agent runs one task, but you can steer an agent to handle multiple tasks together in one context.

00:45:56 • Mo Abualruz

For shared context, you should not have to maintain the context window very much yourself. You cannot directly change the graph here. But if you think the way I build the graph, embed data, and link plans, decisions, and issues is wasteful, with too much data, that can be addressed through code and configuration.

00:46:23 • Mo Abualruz

If configuration is insufficient, that is an issue to address. Especially with smaller models, I try to keep context small and link to files rather than include their whole contents. Some files are huge; they won't fit into a 32K context window. There is some management of that already. We could make the degree of compression configurable, including when to replace content with a file link. For now, the design assumes most people have limited context capacity.

00:47:18 • Mo Abualruz

That is how Fulcrum has been designed.

00:47:32 • Mo Abualruz

Okay.

00:47:34 • Audience (chat question)

You have put a huge amount of work into this local-first project. Where does your motivation come from? What keeps you going? What personal experience led you to start Fulcrum?

00:47:34 • Mo Abualruz

It addresses a need I have. The best solution often starts with your own problem. Once you solve it, try to make the solution more general: how can it help people with similar problems? We have an Arabic saying, “al-hajjah umm al-ikhtiraa,” meaning that necessity is the mother of invention. When you identify a need, don't stop at “These two lines fix it for me.” Go one level further: can I make it robust enough to work for someone else too? As it grows, can I separate it from the project I am working on and make it useful to more people?

00:48:51 • Mo Abualruz

Seeing something work and help you creates a feedback loop. You make work easier and faster, enjoy it more, and reduce token usage. Instead of spending $1,000 a month on OpenAI and Claude inference, you spend $100 here and $100 there, saving $800. Seeing results like that motivates me.

00:49:24 • Mo Abualruz

It may not be the best tool for everyone, but feedback from my own work is important. For any project, look at what you need, what you made, and what value comes back to you. Then ask how to make it more general so that value reaches more people. Perhaps someone will not know your name, but they will know your project benefited their life. That is a way to leave a mark.

00:50:11 • Mo Abualruz

That is where my motivation comes from.

Solar power and the full cost of local computing

00:50:21 • Audience (chat question)

Do you envision planning, orchestration, or collaboration between different users' agents? Also, what would be a good way to set up solar power for local AI?

00:50:21 • Mo Abualruz

Planning and orchestration between different users' agents?

00:50:33 • Mo Abualruz

This is an important context-engineering question. I think we still interact with agents as though they were human. They are not; they are productivity tools, fast typewriters. However smart they are, they remain autocomplete tools. They used to complete a line; now they complete files, pages, and projects.

00:51:12 • Mo Abualruz

Think of AI as tools. People still need to collaborate. If the host and I work together, we need time to talk. Suppose we use the same Fulcrum installation. We can have a call, align our plans, and then ask an agent to check for contradictions, consolidate the plans, and get both our approval. It could do that in five minutes.

00:51:59 • Mo Abualruz

Without first talking to each other, we would not have that workflow. Human interaction remains necessary. But afterward, we need not spend five hours writing down every blocker and conflict. An agent can go through the material and produce a report.

00:52:28 • Mo Abualruz

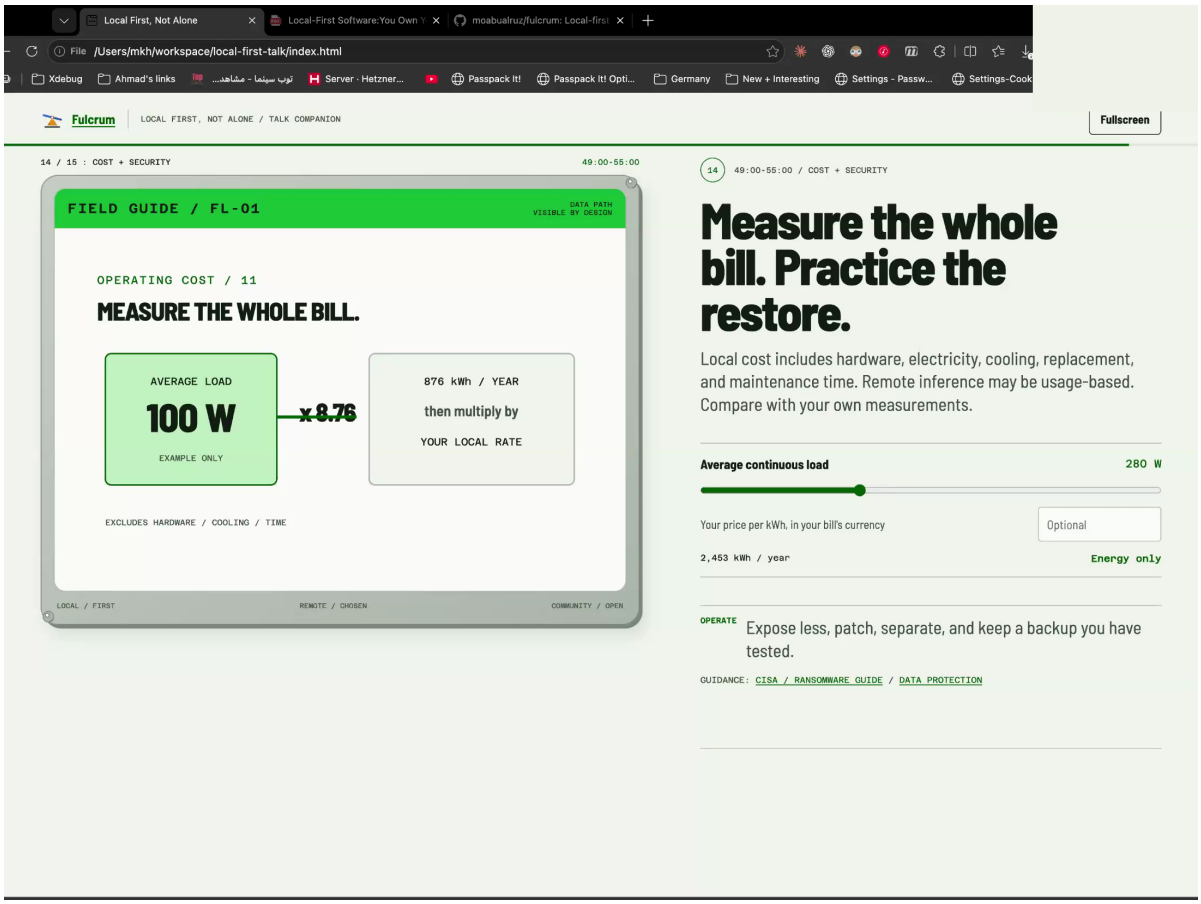
We spend five minutes reviewing it, give the results back to the agent, and it consolidates the plan and runs as many orchestration agents as configured. This is not limited to Fulcrum or a particular harness. Treat humans as humans and AI as tools. I think AI is the best productivity tool since the calendar. Calendars let people coordinate and reserve time; to me, this is the next best thing since that invention.

00:53:14 • Mo Abualruz

The value is huge. I don't like the AI-doomer view; I think it is illogical. Being so afraid of matrix multiplication is too much. People who don't understand the math fear it. AI is a mathematical tool, and it can make you more productive.

00:53:40 • Mo Abualruz

Sometimes it is very smart and sometimes very stupid. Review its work and make sure it does what you want, in the way you want. Then you can be more productive and get the satisfaction of creating value for yourself and others.



Recording frame • 00:55:00

00:54:05 • Mo Abualruz

For local hosting, electricity, cooling, replacement, and maintenance matter. I will go through this quickly because my time is nearly up and I want to leave time for questions. Whether you choose local-first, cloud, or self-hosting, compare your expenses with the value you receive.

00:54:35 • Mo Abualruz

If you want to learn, you can experiment and then tear the setup down. You don't have to keep something running just because you learned from it. Experimentation has costs and benefits. Suppose you want to run a model that is 50 GB: you don't necessarily need to buy an RTX 6000.

00:55:01 • Mo Abualruz

If a friend has one, ask to use it for two hours to run the experiment and learn. If you need it every day, then consider integrating it into your setup. How will spending $15,000 on the card return $15,000 of knowledge, experience, and useful work? Don't buy it simply because it is new or good.

00:55:29 • Mo Abualruz

Get the best value you can. If the cloud gives you the best value for experiments, learning, and projects, use it. Just keep a local backup of your files and knowledge so that a cloud-service failure does not leave you stuck. Here is a quick example: suppose the local machine I run is using 500 watts, and the price per watt is five. I would be paying 21,000 a year for electricity. This is an example of prices in the US. [The spoken electricity-rate unit and calculation are unclear; these figures are retained as spoken.]

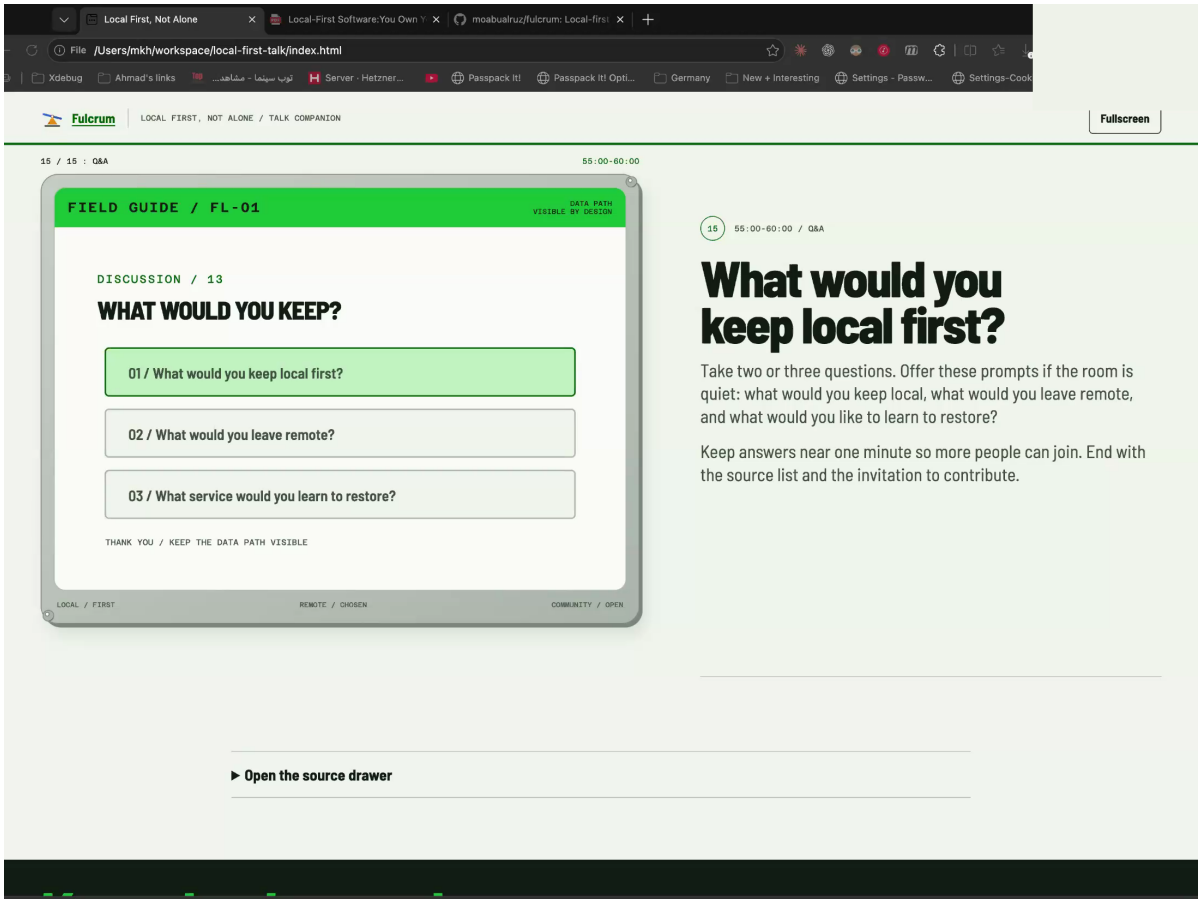
00:56:17 • Mo Abualruz

That is not feasible for me: 21,000 could buy many cloud services for the work I need. I might instead use a smaller machine running on 20 watts to keep files and data organized locally, so I am protected if a cloud provider fails. Balance the value you get from local-first, self-hosting, and the cloud.

00:56:48 • Mo Abualruz

Money does not grow on trees. We need to make sure we get value for what we spend.

Discussion and the Fulcrum demonstration



Recording frame • 00:58:30

00:57:00 • Mo Abualruz

To finish, think about three questions: what will you keep local-first, what will you leave in the cloud, and how will you back up and restore? That is how I approach it.

00:57:26 • Mo Abualruz

Sorry for taking a couple of extra minutes. I can stay for another ten minutes of questions.

00:57:39 • Host

[Translated from Chinese] If you have questions, we have another ten minutes to ask and discuss.

00:58:03 • Host

Which part of the Fulcrum project is most difficult for you? What is the main challenge you are trying to tackle?

00:58:29 • Mo Abualruz

Community-building has been a little slow. But Fulcrum is new, perhaps eight months old, so I can't complain. Market penetration and marketing are difficult even when your project is not open source.

00:58:57 • Mo Abualruz

Technical people need those skills. You don't have to be the best marketer, but you should know how to market yourself, explain your achievements and skills, and communicate the value of the products you work on. It is not deeply technical knowledge.

00:59:29 • Mo Abualruz

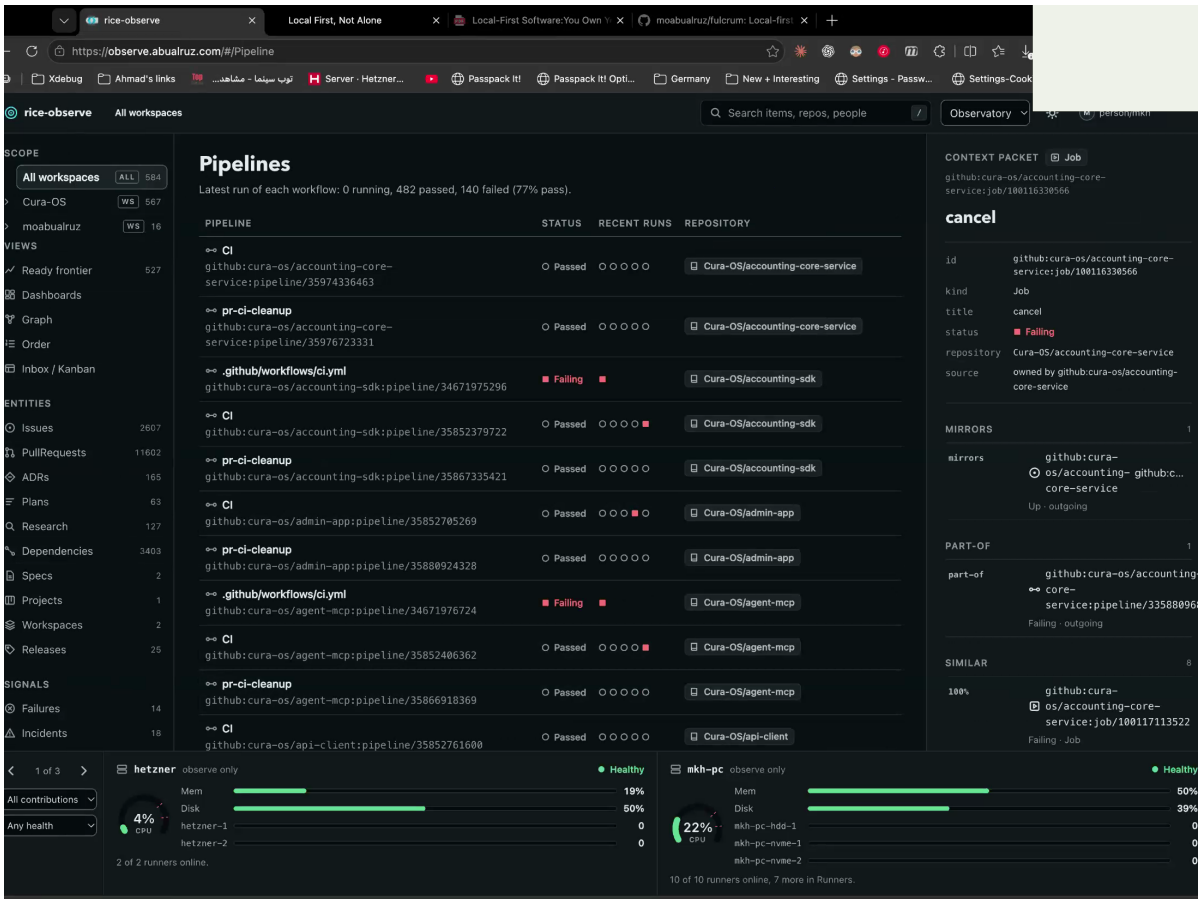
Sadly, I am not a marketing person, and I need to work on that. I can market myself well, but not always my projects. I deliberately called Fulcrum an alpha even though the project is implemented, because without a community I cannot call it battle-tested or say it works well for everyone. I won't claim a beta or final product until I build that community. Six months ago, I left Germany and returned to Jordan. After announcing Fulcrum, I did little marketing: I posted it on X and my website, then moved on. I kept using it without focusing on community-building. That is why today I wanted to discuss not only my project but yours, and the things I am doing wrong in managing Fulcrum.

01:00:49 • Mo Abualruz

Technically, you can use it in a daily setting, not only for development. But I can't think of all the other use cases without a community. Someone may try a scenario and find that one part fails; then we can improve it to support that use. On observability, I am working on a tool that I want to make part of Fulcrum or a sister project.

01:01:34 • Mo Abualruz

Let me find where I put it. Here it is. I need to log in. I have been upgrading security on my local devices, so now I have to enter a one-time code for my own services. That may be excessive for locally hosted services, but you never know who might get your data. Let me bring it up.



Recording frame · 01:02:50

01:02:29 · Mo Abualruz

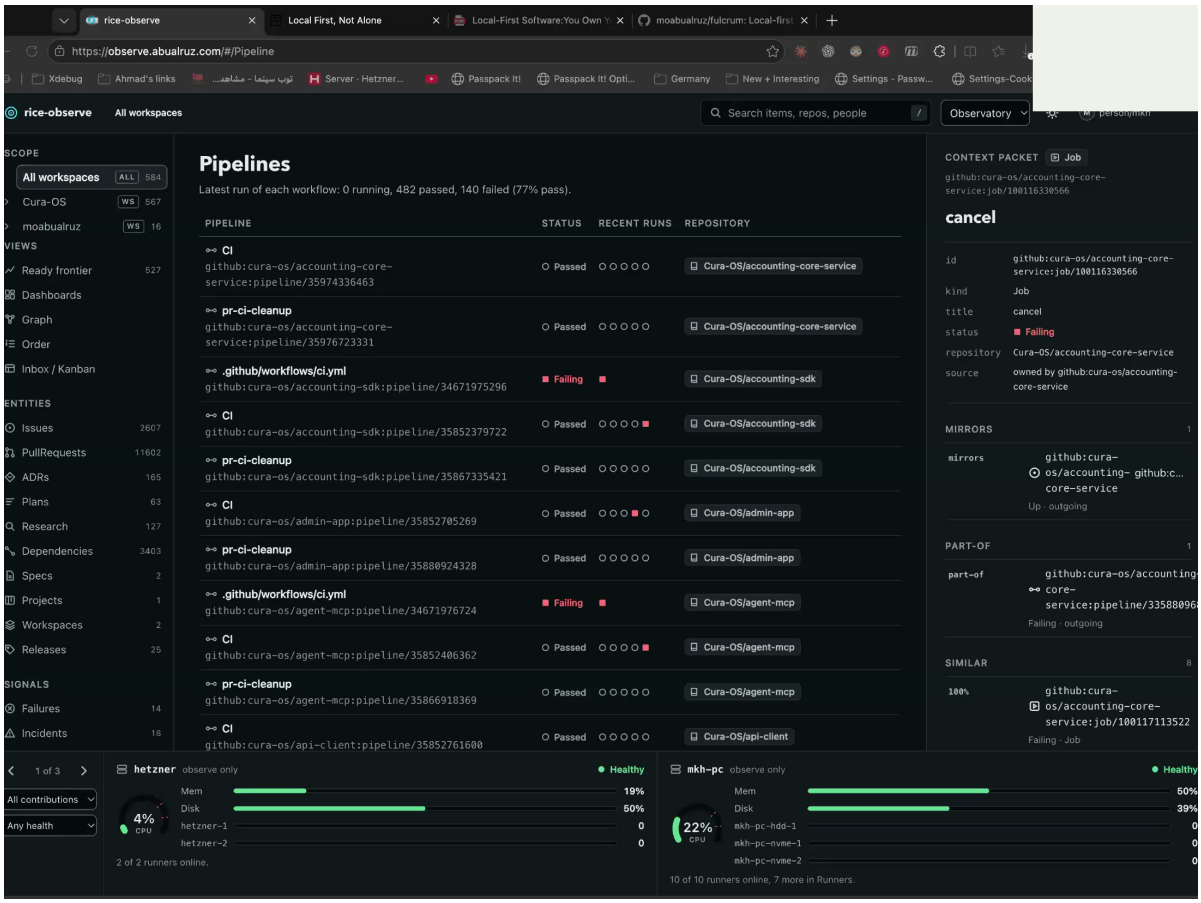
This is rice-observe. It is another platform for the jobs and pipelines I mentioned. I want an integration plugin between Observe and Fulcrum, so there is one place to see what is happening and how everything works.

01:02:57 · Host

[Question unclear; the following answer concerns whether the project has been published.]

01:02:59 · Mo Abualruz

No, it is still in progress and not ready. That is why I haven't published it; the repository is private. I work on something until it stops failing a lot. Then I make it open source, changing the repository from private to public. Right now it still fails often, and I am fixing the configuration.



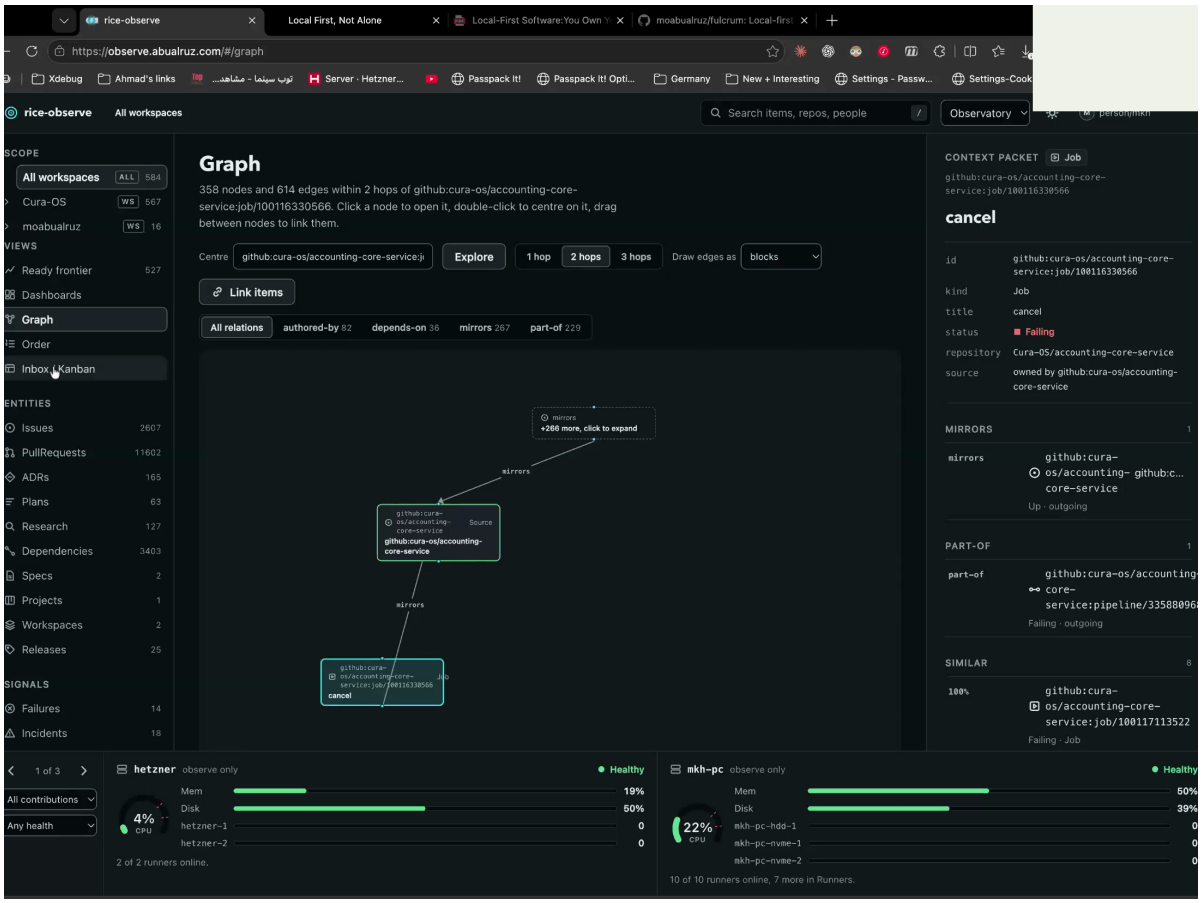
Recording frame · 01:04:00

01:03:32 · Mo Abualruz

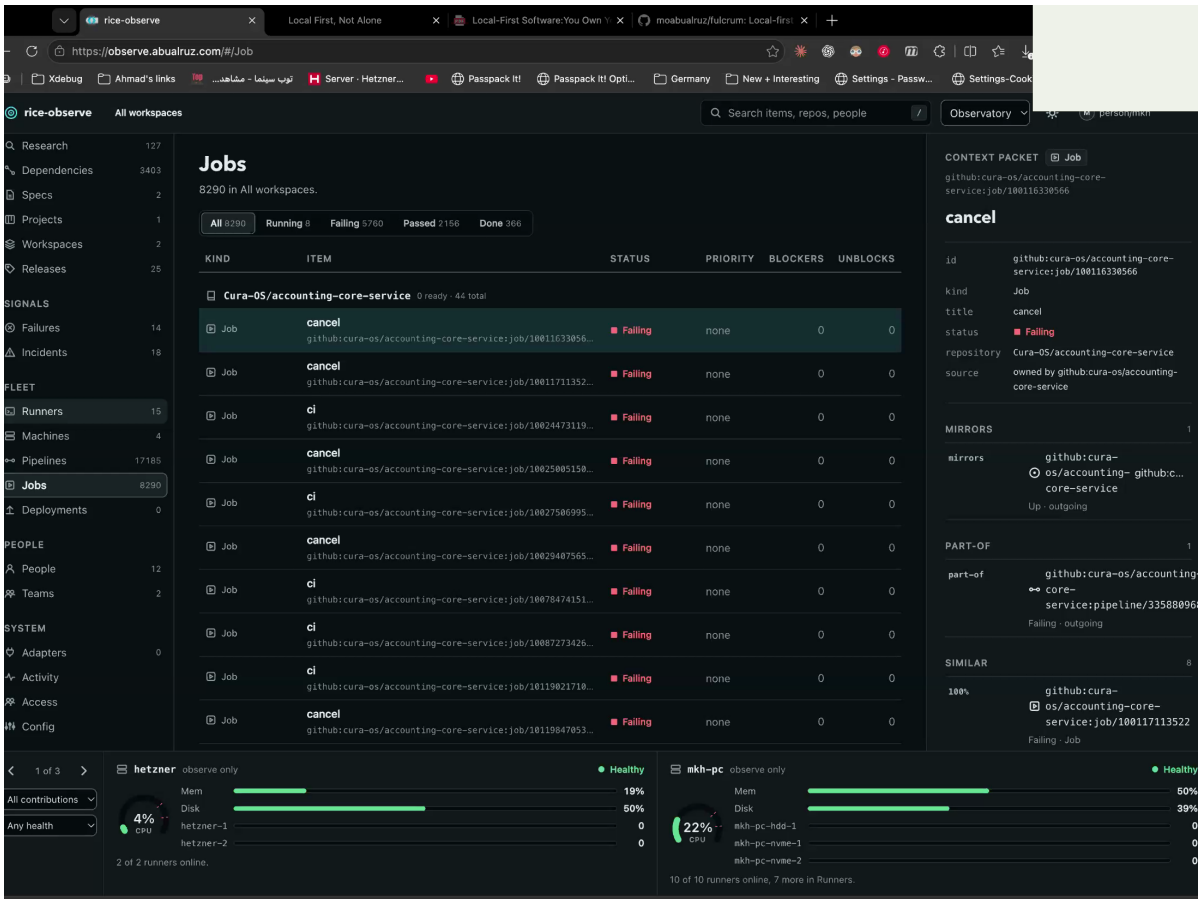
Runners get stuck, queues fill up, and files are not cleaned up after jobs finish. I am fixing many issues. Once it is in good shape, I will announce it as a project of its own, but I also want it integrated with Fulcrum. I want specialized views for ADRs, plans, and runners, while also having one place that integrates everything.

01:04:02 · Mo Abualruz

I don't want to keep jumping between tools. I want to click a link to expand a job, see its details, and understand why it failed. I want a high-level view at all times, with the ability to get down to the smallest details in a couple of clicks. That helps me understand things: I can observe everything quickly, and more detail should never be more than two clicks away. This is rice-observe. I may announce it on X and send the host a link when it is ready. It will enroll all my machines: my Mac, laptop, PC, and a Hetzner node hosting other services. I can check their health and the same graphs Fulcrum builds. Here is an issue I am checking; it uses the same data as Fulcrum. I can also see pull requests running on my projects. It is slow, and I need to improve it. There are many issues, which is why I haven't announced it and the repository is still private. Here are job details: failure after failure, as you can see. I need to fix those before sharing it. That is how I approach the work, and why I made this a separate project.



Recording frame • 01:05:30



Recording frame · 01:06:30

01:06:11 · Mo Abualruz

Some people don't need all of Fulcrum. They don't do much agentic programming, but they have machines and pull requests and do DevOps engineering. Observe is more useful to them. People who need both can use both, and I will build plugins so Fulcrum and Observe work together. Another benefit is avoiding GitHub API limits. When agents repeatedly check things, GitHub may tell you to wait two hours because you have exhausted the quota. I don't want that to prevent me from knowing which PRs are open or which issues I am working on. That is why I sync everything locally instead of continually asking GitHub.

01:07:23 · Host

Yes, that's true.

01:07:26 · Mo Abualruz

If you have more events like this, I would be happy to join. I had fun, and I hope everyone else did too.

01:07:37 · Host

Definitely. Thank you so much.

01:07:42 · Mo Abualruz

Thank you all. I have to go.

01:07:45 · Host

Of course.

01:07:46 • Mo Abualruz

Unfortunately, the other sessions are on weekdays, a little early and during my working hours. I will try to join as an audience member if I can. I love learning from other people's experiences. Thank you all for attending.

01:08:08 • Host

Thank you. Let's keep in touch.

01:08:12 • Mo Abualruz

Thank you. Bye.

01:08:13 • Host

Bye-bye. Thank you.