Explaining Digital Audio to Musicians

Kieran Walsh · November 2026 · kdwalsh.com

Forget musical staves and pitch—why digital audio is really a game of rhythm, rehearsal marks, orbital conductors, and respecting the speed of light.

When we try to explain digital audio to musicians, our first instinct is usually to reach for sheet music—and that is almost always where things go off the rails. As a Royal College of Music–trained horn player who went on to spend a career in audio networking, I've watched it happen from both sides of the music stand.

The classic mistake is pointing to a musical stave and telling a player that the vertical position of a note represents frequency. On paper, it makes intuitive sense: high notes sit at the top of the stave, low notes sit at the bottom. But in digital audio, that analogy is a trap. A single digital sample has no concept of frequency whatsoever. It is simply a discrete snapshot of amplitude at one exact moment in time.

By accidentally equating a single sample with pitch, a well-meaning teacher walks straight into the minefield of Fast Fourier Transforms (FFTs)—the complex maths required to translate time-domain snapshots into frequency-domain spectra. Suddenly, an introductory lesson turns into serious cognitive overload.

To keep things grounded, we have to leave pitch out of it entirely and translate digital audio into something musicians actually do: rhythm, ensemble playing, and reading a score in time.

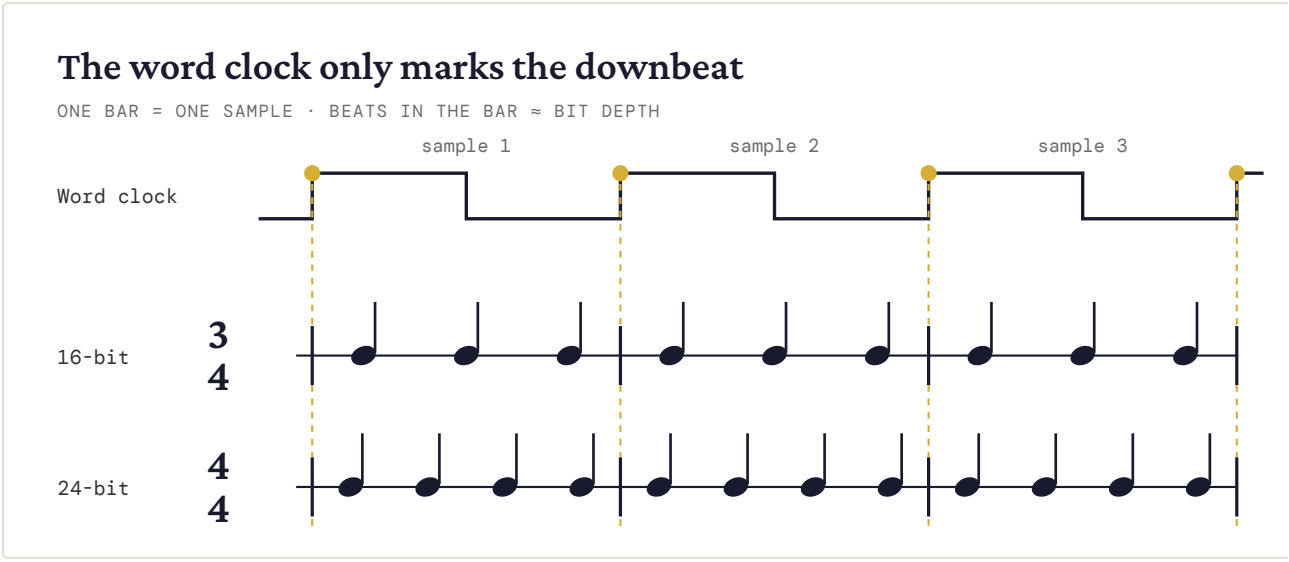
The Anatomy of a Bar: Bit Depth, Sample Rate, and the Word Clock

Start with a single bar of music in simple time—say, 4/4, with four crotchets (or quarter notes) in a bar. In digital audio, that bar represents a single sample.

The internal structure of that bar—how many beats or subdivisions it holds—serves as a useful structural analogy for bit depth. While time signatures and binary word lengths aren't literally the same thing, thinking in subdivisions gives us a shared vocabulary for how the data is organised inside each packet. Standard professional audio uses 24-bit samples, while older CDs used shorter bars, like 16-bit (a bit like dropping down to 3/4 time—though it's safe to say CDs don't necessarily make better waltzes than modern professional audio). 32-bit audio is a different story altogether, functioning much more like compound time, with extra subdivisions to accommodate far greater dynamic complexity. What those extra subdivisions really buy you is quieter silence: each additional bit lowers the noise floor by about 6 dB, which is why 24-bit offers roughly 144 dB of theoretical dynamic range against 16-bit's 96 dB.

Next, we bring in a conductor to establish the sample rate, which determines how many bars pass in a given unit of time. If the conductor sets an Allegro tempo of 120 beats per minute, they get through a 4/4 bar every two seconds. Digital audio works on the exact same principle, just at a vastly higher speed—typically 48,000 bars every second (48 kHz).

This is also where one of professional audio's most persistent myths creeps in: the belief that a word clock somehow provides superior timing accuracy or magic resolution. In reality, a word clock is nothing more than the conductor's downbeat. It simply sends a pulse to mark the start of the bar—the beginning of the sample frame—and tells you nothing whatsoever about how many beats or subdivisions sit inside it. Think of the famous ballroom scene in Mozart's *Don Giovanni*, where three separate orchestras play simultaneously in three different time signatures (3/4, 2/4, and 3/8), all locked together at the barline. A word clock works the exact same way: you can have a 3/4 waltz and a 4/4 bar running side by side, and the word clock merely ticks on beat one of each bar to keep their downbeats aligned, while the 3/4 bar naturally spaces its three beats proportionally slower across that same window of time.



Different beats inside, same barline: at 48 kHz the word clock ticks 48,000 times a second and says nothing about how the bar is subdivided.

From Conductor's Score to Single Wire

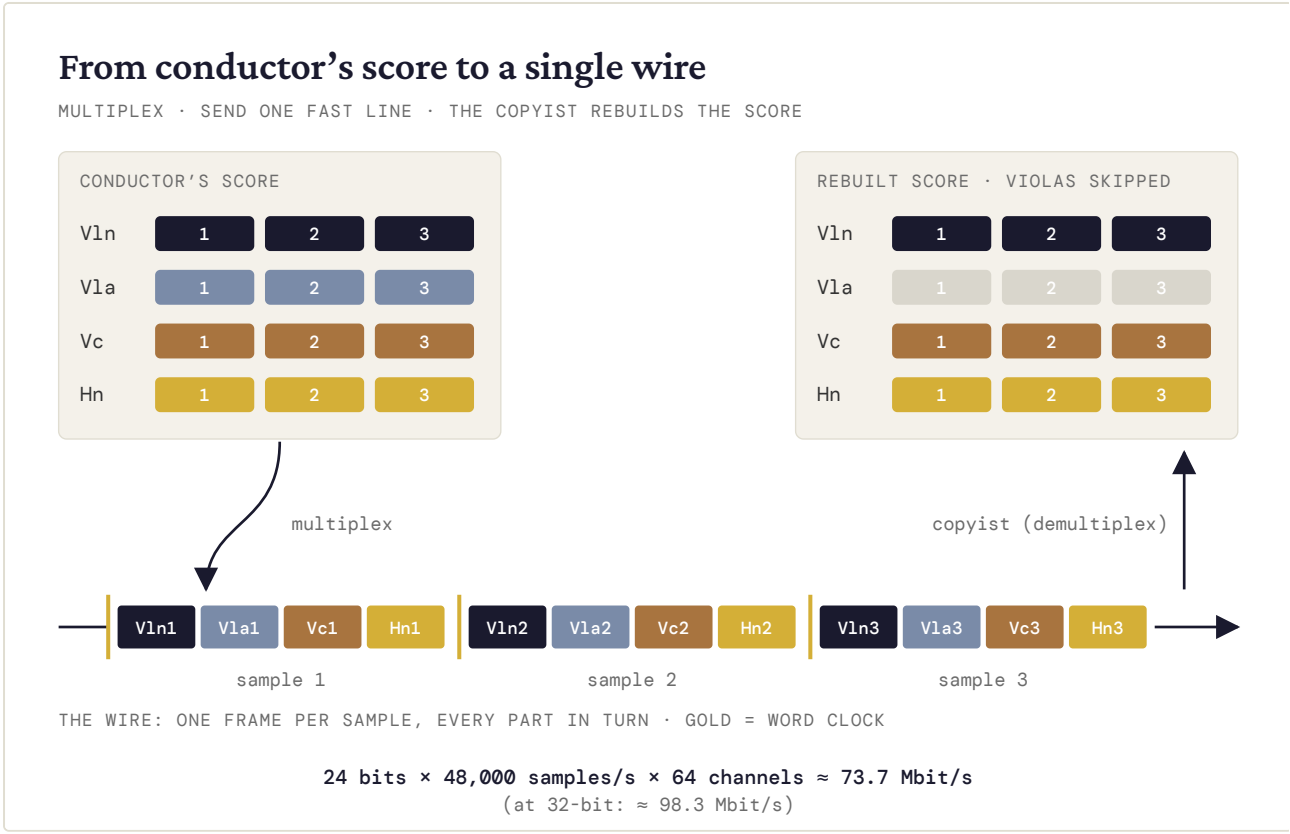
When we move from a single instrument to a full ensemble or multi-microphone setup, we put everything onto one large conductor's score. Looking vertically down the page, we can observe every single part—or audio channel—happening simultaneously. However, when we want to send all of that audio down a single wire, we have to multiplex the signals together.

To do this, we take all those vertical parts and lay them out end-to-end into one continuous, horizontal line. To figure out how fast that single line has to move, we multiply everything together:

> **Data rate = Bit depth (bar size) × Sample rate (tempo) × Channels (instruments)**

For example, running 32-bit audio at 48 kHz across 64 channels requires a data rate of roughly 98 Mbit/s (or about 74 Mbit/s at 24-bit).

To a musician, squeezing a 64-piece orchestra into one impossibly fast, single-line melody makes no sense—trying to read and extract your individual notes from that stream would be completely bewildering. Yet, that is precisely what digital audio matrices and cables do. When the signal reaches the other end, the receiving system acts like a high-speed copyist, chopping the individual parts back out of that long, frantic line and reassembling them into a readable score so every channel can be heard cleanly again.



Every part, sample by sample, in one fast line — and the copyist only transcribes the parts it needs.

Throwing Out the Old Copyist Rulebook

In large-scale professional setups, however, we aren't just dealing with 64 parts—we are often dealing with 1,000 parts or more. In the old days, a copyist would take that incoming stream of notes, meticulously transcribe every single bar of every single part into a giant holding area—what we call a buffer—and once the whole score was assembled in memory, pass it off to the "printer" to be played back.

Modern, high-speed audio engineering throws that old rulebook out the window:

- **Selective listening:** First, if we don't care what the violas are doing—which, as any orchestral player will tell you, is generally quite a good idea—we don't waste time transcribing them. We simply ignore their notes as they fly past in the stream, saving valuable processing power.
- **Zero waiting:** Second, we don't assemble a massive score just to send it to a printer; we deal with the notes in real time, processing them the instant they arrive.

- **Independent buffers:** Most importantly, we don't force every musician to wait on the same giant holding buffer. Most of the orchestra has to stay tightly locked to the conductor's baton, so their parts go into a strictly synchronised buffer so everything stays perfectly in time.

Imagine, though, that we have an off-stage viola—a scary thought, especially with a microphone in front of it. Because they are far away, their sound is going to arrive late anyway, whether they're in the pit or out in the corridor. Rather than holding up the entire orchestra waiting for that distant signal, we route it into a completely independent buffer. When the off-stage viola finally shows up, we simply deal with it as fast as it arrives and get on with it.

My approach breaks sharply from industry consensus—a theme I explore further in my piece Beyond Compliance – the ideal. Most conventional systems still rely on massive holding buffers, which forces the central brain—the CPU—to work far too hard copying and shuffling data around. By using a zero-copy architecture, the audio stays right where it is in ultra-fast, on-chip RAM. Instead of reading every individual note, the CPU only tracks pointers—functioning much like a conductor who only needs a list of rehearsal numbers to put every musician in the right place without ever getting bogged down in note-by-note detail.

Locking the Score in Transit

In a modern world of deepfakes and security threats—especially when dealing with live speeches or sensitive broadcasts—we also have to protect the audio in transit. To stop anyone tapping the wire from listening in, we apply encryption. Using a secret cipher (or key), we scramble the musical stream into ciphertext, making it computationally infeasible for an eavesdropper to reconstruct the score without the key. When the scrambled stream arrives at the other end, we apply the matching cipher using a straightforward logical operation called an XOR, which instantly snaps the scrambled notes back into readable plaintext.

In the old days, doing that kind of heavy cryptographic maths meant halting everything inside that giant holding buffer, choking the CPU, and introducing unacceptable delays. Today, we offload that work to dedicated cryptographic accelerators built right into the silicon: an entirely separate, specialised brain whose only job is to unscramble the ciphertext on the fly.

But encryption alone is only half the story. While the exact mechanics of the cryptographic primitive behind my system are subject to the processes of the UK Intellectual Property Office—and remain tightly under wraps—the principle is straightforward: alongside the encryption keys, it binds identity and provenance directly to every single sample as it flies past. Think of it like a parcel delivery: encryption locks the box, while this method ensures that a parcel sent by Royal Mail only ever shows up in a genuine Royal Mail van. Every single bar of audio carries cryptographic proof of exactly where it came from, which is what stops spoofing and deepfakes cold.

Putting the Conductor in Orbit

Making that provenance work globally comes down to extreme precision in time. When a stream is moving at 98 Mbit/s, you are operating in microseconds—and to measure that accurately, you have to look three orders of magnitude deeper, down to nanoseconds (1,000 nanoseconds to a microsecond).

To lock onto an originating timestamp within ±20 nanoseconds anywhere on the planet, we take our downbeat from a conductor orbiting roughly 20,000 kilometres (12,500 miles) above our heads: GPS. Frankly, launching the conductor 20,000 kilometres into orbit seems like a wonderful idea to most orchestra members—though in this case, it is also an engineering necessity.

Of course, listening to just one satellite isn't enough; atmospheric interference and orbital drift mean any single timing source has slight errors. To find the exact, indisputable truth on the ground, the receiver has to read multiple satellites simultaneously, strip out atmospheric distortion, and reconcile their signals through cross-correlation and weighted estimation—checking each incoming timing pulse against the others, accounting for orbital mechanics, and verifying that every satellite is telling the truth. By resolving those orbital downbeats into a single, rock-solid point of truth (a timing discipline closely related to my work on An Interesting Open Source PTP Grand Master for Proxmox), we anchor every single sample to an indisputable moment in time and space.

The Paradox of Performance: Why Provenance Beats Secrecy

All of this high-speed engineering—the multiplexers, the zero-copy buffers, the orbital clocks—exists for one ultimate purpose: creating a musical performance. And the entire point of a performance is sharing it.

In classic cryptography, we always talk about Alice sending a secret message to Bob. That model works brilliantly when you want total secrecy, assuming nobody can see Bob's screen. In audio, however, that model hits a brick wall—specifically, the laws of physics. The moment Bob's loudspeaker pushes air into an acoustically coupled room, the secret is out. Anyone in the space can hear it. And from a musical perspective, that is the whole point. We aren't trying to hide the music from the audience; we are trying to share a coherent mix with them.

That coherence is why, despite having the capacity to move a thousand isolated channels, we still have to treat a symphony orchestra as a single instrument. People often assume modern orchestral recording requires an army of microphones, yet for years Abbey Road Studio One—built to hold a full symphony orchestra—was specified with 48 channels of remote mic preamps—and even that is far more than you would typically use on a real orchestra in a great-sounding hall. Horses for courses, of course; innovative recording techniques are always welcome, and tradition is never an excuse to avoid progress—Abbey Road itself reopened Studio One in 2025 with an 84-channel console and 7.1.4 immersive monitoring, and quite right too. Still, there is a specific, blended sound that the public imagination expects when they hear an orchestra, and good production means honouring that coherence rather than fracturing it into isolated fragments.

This brings us to a fascinating inversion in how we think about audio security. Once the music enters the room—or the broadcast—confidentiality stops being the main event; attestation to the source takes over.

For decades, authentication has been the poor relation in the cryptographic family. The James Bond allure of scrambled signals and secret codes is far more glamorous than the seemingly mundane question of who actually said it. Yet in an era of synthetic media and real-time deepfakes, proving the origin of a sound becomes just as vital as protecting its content—and often even more so.

The Remote Orchestra Myth—and Where Low Latency Truly Shines

Finally, we have to confront a popular fancy in modern audio: the idea of the distributed remote orchestra, where the strings sit in London, the brass in Edinburgh, and the percussion in Amsterdam, all playing together in real time as a single, tightly locked ensemble.

Whenever someone proposes this as a replacement for a live hall, they run headfirst into an unforgiving adversary: the laws of physics. Signals take time to propagate. Even inside a fibre-optic cable, light doesn't travel at its full vacuum speed of roughly 3×10^8 m/s (300,000 kilometres or 186,000 miles per second); glass slows it down to around two-thirds of that speed—roughly 2×10^8 m/s (200,000 kilometres or 124,000 miles per second). Across a room, that feels instantaneous. Across countries and oceans, distance becomes an insurmountable wall.

You will sometimes see research quoting an "ensemble performance threshold" of around 20–25 milliseconds, below which musicians can supposedly play together as if they shared a room. Try playing funk at 20 milliseconds. That is roughly seven metres of air, the drummer and bassist on opposite sides of a big stage, and the groove simply falls apart. And that budget is spent long before a note is played: real fibre routes are never straight lines, and every switch, buffer, converter and monitor mix adds its share.

But the numbers only tell part of the story, and they vary enormously from player to player. A rhythm section doesn't lock together by listening alone. It locks together because each player predicts the others: the breath before an entry, the lift of a bow, the drummer's shoulder dropping into the fill. That shared anticipation is how an orchestra organically keeps time with itself. Much of it lives in micro-gestures and micro-expressions that even the best camera on the planet never captures, which is exactly why video conferencing has never truly replaced being in the same room. Anyone with a career in ensemble playing is finely tuned to pick up these tiny signals. How much they really carry deserves research of its own, but that lies beyond the scope of this article.

I was once asked about synchronising an audio source in Amsterdam with a source in New York, with the strict demand that it happen within one millisecond. Unfortunately, even at the speed of light in a vacuum, crossing the Atlantic takes roughly 20 milliseconds—and in real-world fibre-optic cables, closer to 30 or 40 milliseconds one way. Unless someone has discovered a stable wormhole between Europe and North America (which we would probably notice, given its catastrophic effects on our existence)—and quantum entanglement notwithstanding, which still refuses to carry actual information faster than light—achieving a 1-millisecond transatlantic sync isn't an engineering challenge; it's a violation of physics.

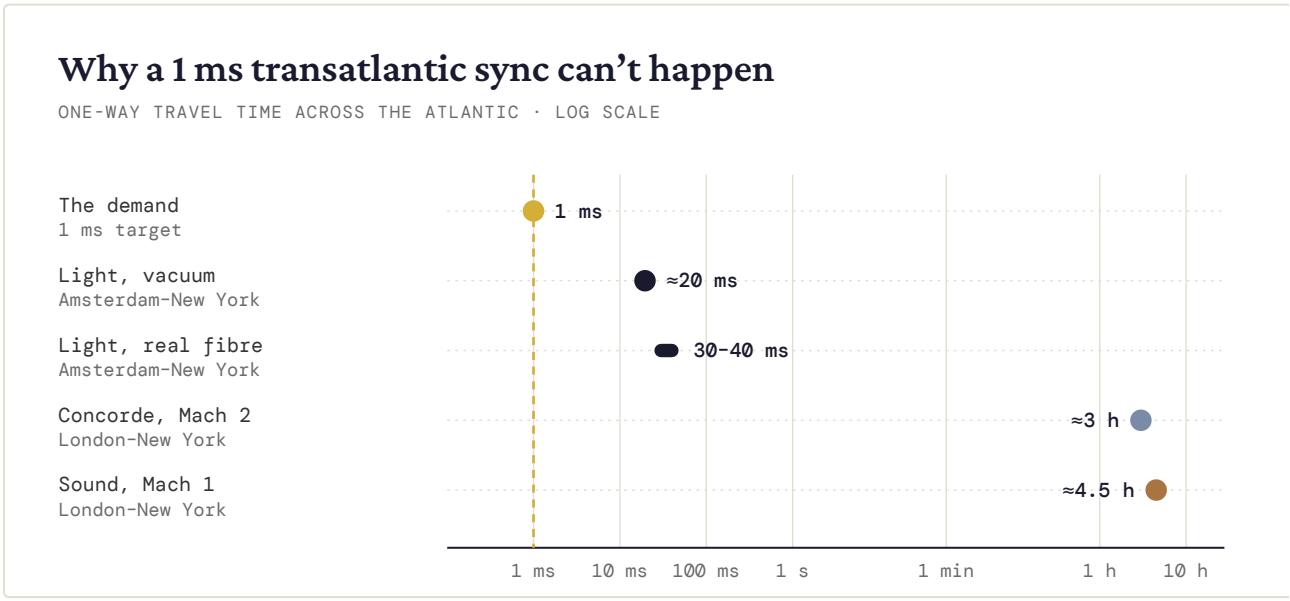
To put that into perspective, I asked the client if they remembered Concorde.

“Of course,” they said.

“And how long did it take Concorde to fly from London to New York?”

“About three hours—but that’s an aeroplane, not sound.”

Exactly. Concorde flew at twice the speed of sound, and it still took around three hours to cross the Atlantic. An actual acoustic wave travelling through the air at Mach 1 takes roughly four and a half hours to make that journey. Fibre optics are unimaginably faster than sound, yet even light cannot compress an ocean into a millisecond.



One-way travel time across the Atlantic. The 1 ms demand is twenty times faster than light in a vacuum allows.

That doesn’t mean networked performance has no value—far from it. When musicians are isolated from one another, distributed performance is a wonderful tool. Pioneering work like Stanford University’s JackTrip and virtual choirs has proven just how vital networked music-making can be for human connection, giving isolated performers a genuine psychological and spiritual lift when they cannot be together. However, we have to distinguish between a joyful lifeline during isolation and the acoustic reality of a major ensemble. For a professional symphony orchestra chasing true acoustic cohesion, no amount of networking wizardry beats booking a few coaches, reserving some hotel rooms, and putting the musicians in the same space.

Once we stop demanding the impossible from physics, the actual edge cases where transit delay causes a real problem turn out to be vanishingly small compared to the vast range of workflows where this architecture shines:

- **Remote studio tracking and film post-production:** In commercial music, a remote artist can easily lay down overdubs in time with a pre-delivered guide track, ready to be blended in the final mix. Similarly, in film Automatic Dialogue Replacement (ADR), an actor can re-voice their lines from across

the globe because the track is resynchronised when the final cut is assembled. Even better, when every sample carries a precise timestamp, we can measure the exact transit delay through the pipeline itself and compensate for it automatically—a deterministic timing problem I find personally fascinating.

- **Remote sports broadcasting:** Having worked on major events at Wembley Stadium syndicated around the world, I've seen firsthand how seamlessly a remote production workflow operates when the control room sits on another continent. You give direction to the crew on the ground, they deliver the shot, and you see the result a few tens of milliseconds later. That tiny propagation delay is completely irrelevant—especially when the finished broadcast is about to be fired into space, bounced off a satellite, and beamed down to viewers' living rooms (which, spoiler alert, takes vastly longer than the entire production chain combined).
- **High-stakes testimony and field operations:** Consider a remote witness giving confidential evidence to a judge in a courtroom, an emergency responder feeding critical intelligence back to a command centre, or a corporate team operating across borders. In these environments, decisions aren't made on the downbeat of a conductor's baton—they are made over seconds, minutes, or hours, often after information is transcribed and reviewed. A few dozen milliseconds of fibre delay changes nothing about the substance of the message. What matters is having the latest, cryptographically verified truth delivered as fast as the laws of physics allow, with complete confidentiality and unshakeable proof of who is speaking.

By keeping our focus grounded in physical reality—treating samples as discrete moments in time, anchoring them with orbital precision, and knowing when to trust the acoustics of a room versus the speed of a wire—we unlock solutions for problems that once seemed impossible: real human music when we gather together, and rock-solid, authenticated communication whenever we are apart.

© 2026 Kieran Walsh. Online version with links: https://kdwalsh.com/articles/explaining-digital-audio-to-musicians/
Contact: website@kdwalsh.com · +44 1273 117672