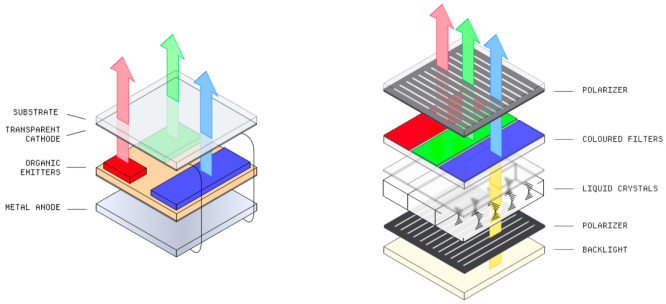
3589 WORDS | DAN HOLLICK

How does a screen work?

From electron guns to tiny electric crystals - digital displays have always been the unsung hero of computing.



OLED PIXEL LCD PIXEL

I understand the irony of starting this book about software, talking about hardware. But there isn't a more under appreciated technology in modern computing than digital displays. In fact, modern computing just isn't possible without them. If we developed the transistor before the CRT, I'm not sure you would even be reading this right now.

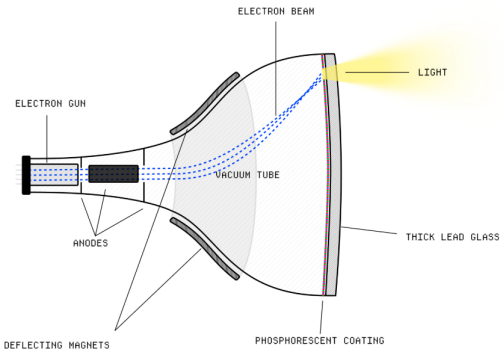
The reason it's so under appreciated is because most people have no idea how a screen works. Any time you see a pixel light up, you are witnessing actual witchcraft before your eyes - light bending through electric crystals just so you can read a tweet in bed.

A brief history of digital displays

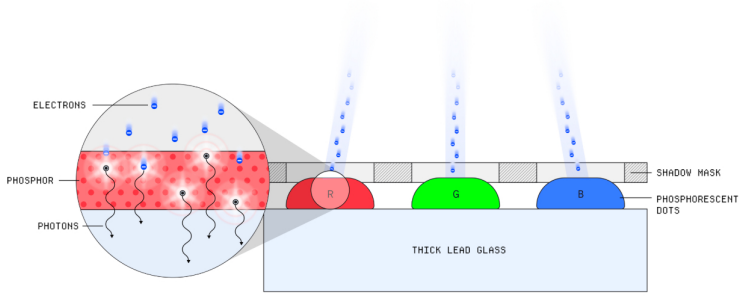
Before we can understand where we are, it's helpful to understand where we started.

The first practical electronic display was the CRT, or Cathode Ray Tube, which has been around in some form since 1897 and became a feasible household appliance in the 1930s. These early Television sets were monochrome, round, and dim, and it wasn't until the 1950s that the large rectangular color units we recognize were introduced.

Although the way they work is conceptually simple, it's remarkable that it was possible to shrink down into something we all had in our homes. It is essentially a giant glass vacuum tube, inside which is an electron gun that fires a beam of electrons towards the front screen.

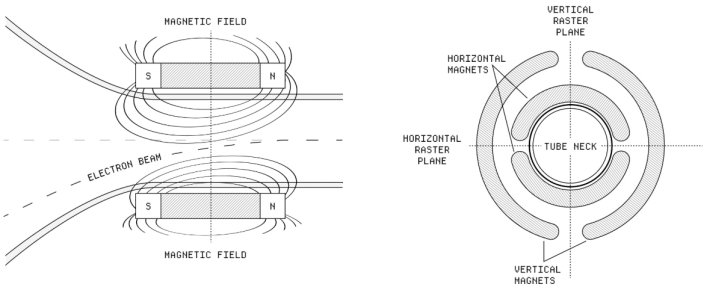


The screen is coated in tiny phosphorescent dots which emit red, green, or blue light when hit by the electron beam. Altering the intensity of the beam affects the intensity of the light emitted by each sub-pixel and in turn allowing different colors to be produced.



It sounds almost like science fiction but for half a century we all stared down the barrel of an electron gun for hours at a time.

Around the neck of the vacuum tube is a set of magnetic coils used to deflect the beam to different areas of the front screen. Changing the direction and magnitude of the current flowing through the coils manipulates the magnetic field, allowing precise control over the path of the electron beam.



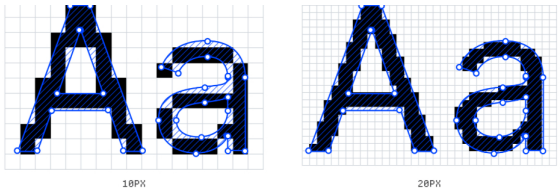
The front screen is usually curved slightly to prevent the geometric distortion that can occur when the electron beam hits the phosphors at an acute angle. The beam paints the image, one pixel at a time, in a raster scan pattern, typically from left to right and top to bottom, 60 times a second.

Unsurprisingly, CRTs had their disadvantages. They are incredibly heavy, thanks to the thickness of glass needed to create a stable vacuum tube for the electrons to travel undeflected by air particles. This also set a limit on how large the display could practically be, while the distance required between the screen and electron gun set a limit on how small they could be made too. Not to mention that they are fairly power hungry - turns out firing electrons out of a gun and deflecting them with electromagnets is not particularly efficient.

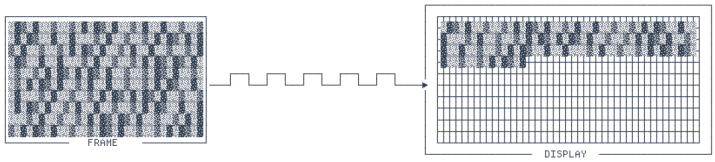
It's fair to say that early computing doesn't happen without CRTs but I think it's also fair to say that portable computing doesn't happen with them. The industry moved to develop thinner and lighter flat panel displays - settling on LCD and later OLED after a brief flirtation with Plasma displays.

Why pixels?

With hindsight, this might sound like an odd question but using pixels wasn't always that obvious for computing. Pixels aren't perfect - dividing an image up into a grid means you run into problems as soon as something doesn't fall neatly onto the grid. You need to rasterise vectors, like fonts, which results in jagged aliasing artifacts in curves and diagonal lines, especially on those early low resolution displays.



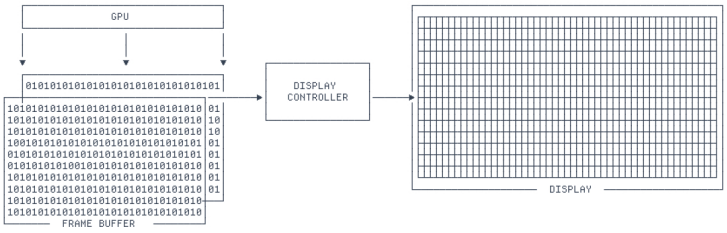
The choice to use the pixel and raster approach has a lot to do with the fact that CRTs already worked this way. Analogue television signals were transmitted in sequence - meaning that the image was broken down from left-to-right and top-to-bottom and then reconstructed in sequence by the CRT.



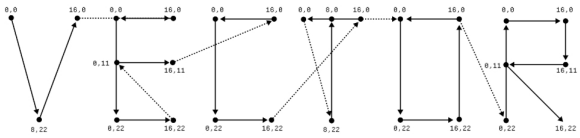
This only really works because the phosphors keep emitting light for a short time after being hit by the electron beam, not quite long enough to paint the rest of the image before going black again - but long enough to trick our eyes into thinking they see a stable image. The advantage of this approach is that it uses very little bandwidth and the hardware inside the CRT only needs to be powerful enough to decode one part of the image at a time.

Early computers adopted this, breaking the image into a discrete grid of pixel values and sending those values, in sequence, to the CRT. But this requires a fair bit of CPU capacity and so gradually dedicated video hardware was developed to offload some of this. Eventually, when memory prices made it practical, the color value of every pixel of the screen was stored in something called a framebuffer.

This is a dedicated block of RAM that contains a two dimensional array, or matrix, of the red, green, and blue values for each pixel. The system writes to the framebuffer when there has been a change to the interface being displayed, only needing to update the parts that are different. The display controller reads from the framebuffer at the refresh rate of the display and so it's up to the system to make sure that the framebuffer is updated in time. ⁂



But before the framebuffer approach became feasible, there was another display technology. Vector displays (or Calligraphic displays) offered a completely different approach - instead of scanning a pixel grid line by line, the electron beam is steered between the co-ordinates of a vector. You can sort of think of this like drawing an SVG, where the image data is a set of instructions.



The upside of these displays is that they used less memory, only needing to store the vector information for anything currently on screen. Some of them used phosphor coatings that glowed for minutes and therefore needed no dedicated memory at all. Although, they still had

glowed for infinites and therefore needed no dedicated memory at all. Although, they still had discrete phosphor dots you could argue are pixels, they didn't suffer from aliasing or pixelation - they were mostly monochrome and so the phosphor dots could be tightly packed.

The downside is that they could only draw lines and no solid shapes, which meant the style of text you could render was limited. They also had their own technical complications, like a slow refresh rate due to moving the beam arbitrarily across the screen. Thus they were mainly used for specific applications like fighter pilot HUDs, radars, oscilloscopes, and most memorably the Asteroid arcade game.

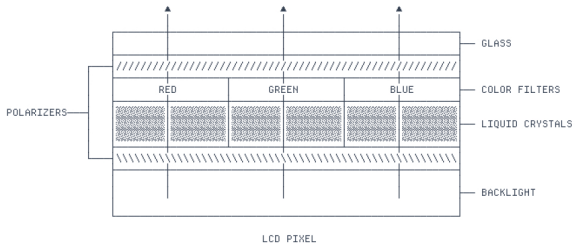
Eventually, memory became cheaper and graphics processing power increased making raster displays the dominant technology. Even though modern displays don't paint the image line-by-line, we still use the framebuffer approach and we have gotten around the loss of fidelity in curves by increasing resolutions and making pixels smaller and smaller.

Modern displays

The modern display war is being fought between two types of technology, OLED and LCD. But it's actually a proxy war being waged by two completely different approaches to displaying an image, transmissive vs self-emissive. Each side is racing to produce the holy-grail of displays: pure blacks while still being extremely bright, fast refresh rate and perfect color accuracy. Oh, and don't forget being cheap to produce with low power consumption.

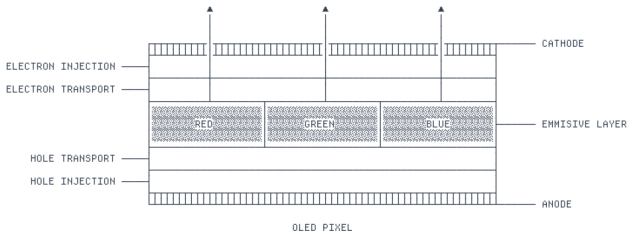
Modern displays are different from those CRTs in one particularly crucial way. They light up each pixel simultaneously, refreshing the entire display at once. They can do this because each pixel is controlled independently, instead of being controlled by a single source.

A transmissive display, like an LCD, is typically one that uses some sort of backlight and the sub-pixels alter the intensity and color of the light as it passes through them. Their weakness is usually light bleeding from the backlight resulting in poorer contrast and narrower viewing angles. But they excel in brightness, cost and lifespan.



LCD PIXEL

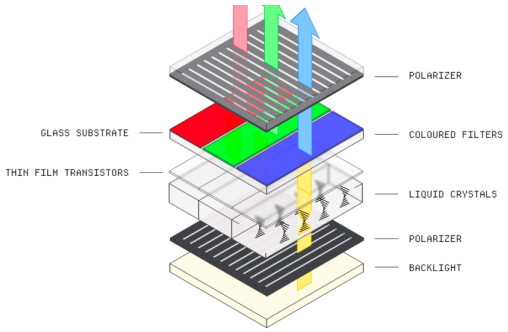
A self-emissive display, like an OLED, typically has no backlight and instead the sub-pixels emit their own light. Their weakness is usually brightness and lifespan, as they can burn-in over time, but they are energy efficient, responsive, and able to produce pure black by turning off individual pixels.



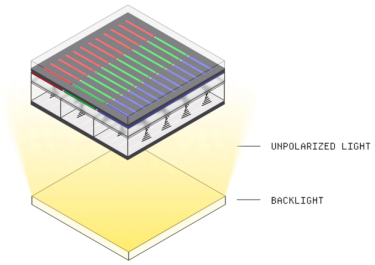
OLED PIXEL

How an LCD works

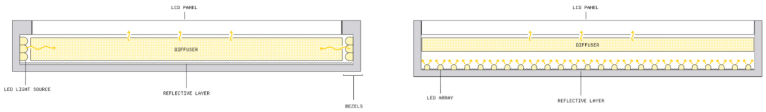
Liquid Crystal Displays are by far and away the most popular flat-panel display technology - it's almost a certainty that you use an LCD daily in some application. The fact that they ever made it out of the research lab and into our homes is astonishing to me.



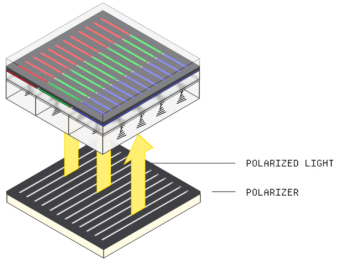
They are unique in being transmissive, whereas CRTs, Plasma, and OLEDs are all self-emissive. Each LCD pixel consists of a layered stack with a backlight at the bottom - each layer conditions the light in some way as it passes from the backlight to your eye. Lets go through it layer by layer, starting with the backlight.



Up until a few years ago, most LCD backlights consisted of some LEDs around the edge of the display that would shine into a diffuser. The diffuser helps even out that light and re-directed it up into the panel. This is why older LCDs have such thick bezels, they needed to house the backlight LEDs.



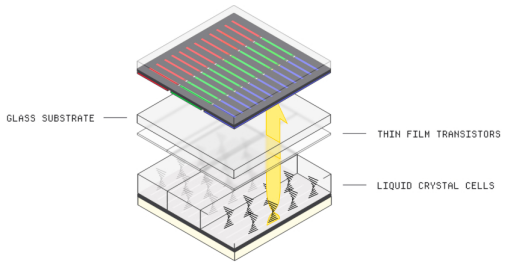
Because of advances in LEDs, these days most backlights are an array of LEDs behind the panel which creates a much more even spread of light throughout the panel. The light then passes from the backlight into the first polarizer.



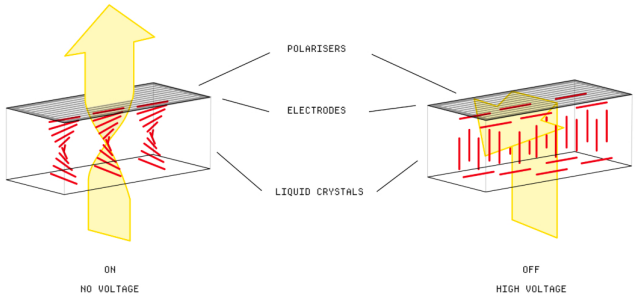
Light is weird - it's both a wave and particle but the waves fan out perpendicular to the direction of travel. A polarizer is essentially a sort of grate that only allows waves in a particular orientation to pass through it.

This is an important step because we need the light that passes through to liquid crystal layer to have a uniform orientation. The crystals will then sort of bend the light so it's either blocked

or let through by the second polarizer.

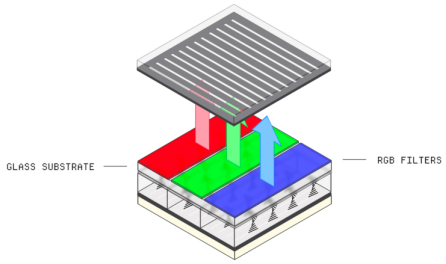


Liquid crystals are straight up magic. There are different types, but in essence they are transparent crystals that change their structure when exposed to an electric field. In their default state, they are arranged in a kind of helix structure that allows light coming from the first polarizer to pass through. When we apply a voltage to transparent electrodes either side of them, the electric field causes their natural helix structure to untwist and align more vertically, blocking the light.

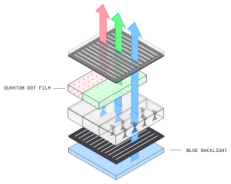


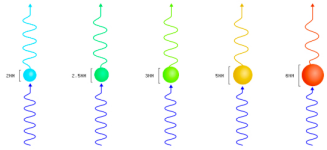
This helix allows light through while changing its polarization - sort of rotating it 90 degrees - allowing the light to pass through the second polarizer. We can modulate the voltage to control the degree of polarization that passes to the next layer and therefore its intensity. Each sub-pixel has its own individual cell of liquid crystals, allowing us to control the intensity of each red, green and blue component separately.

Next, each sub-pixel has an either red, green, or blue filter. This takes the usually white light and blocks out other frequencies allowing the components to be mixed into the desired color.

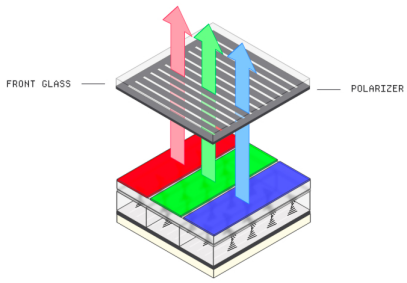


Newer displays, especially those made by Samsung, replace these filters with a quantum dot layer. Quantum dots are microscopic semiconductor nanocrystals that absorb light and re-emit it as a specific color determined by its size. In these displays the backlight is usually blue, because blue light has a shorter wavelength and higher energy, and then that light is re-emitted as green or red by the quantum dot layer.



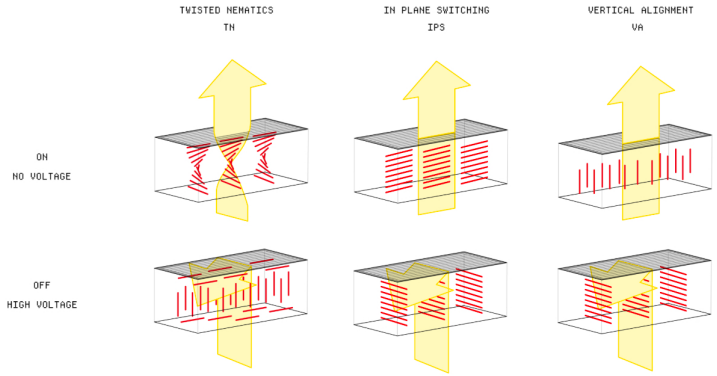


After the color filters, any light that makes it through passes through a second polarizer, orientated 90 degrees from the first one. This allows any light that's been correctly polarized by the crystals to pass through and blocks any light that isn't.



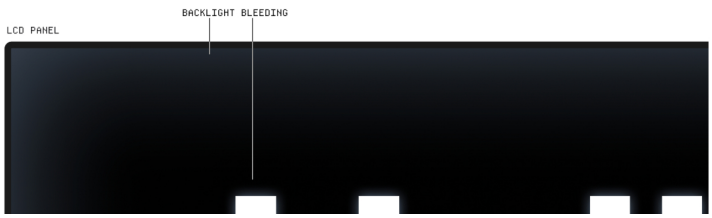
There's a bunch of crazy things required to make this work, like TFTs, or Thin Film Transistors, which are tiny switches manufactured directly onto the glass substrate - one for each sub-pixel. Bear in mind, there can be tens of millions of pixels in a display.

The type of liquid crystals I've described are TN (Twisted Nematic) but there are newer technologies like IPS (In Plane Switching) and VA (Vertical Alignment). The primary differences are the orientation of the crystal structure - the goal is to overcome some of the drawbacks of TN crystals which aren't very good at blocking light in the off position and have relatively poor color accuracy.

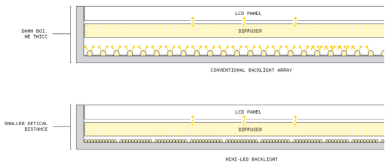


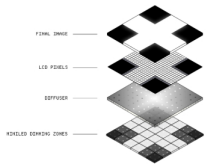
	TN	IPS	VA
Structure	Crystals twist 90°	Crystals rotate parallel to screen	Crystals align vertically, tilt
Viewing Angles	Narrowest (~170°/160°)	Widest (~178°/178°)	Wider than TN, narrower than IPS
Contrast Ratio	Lowest (~600:1 - 1200:1)	Moderate (~700:1 - 1500:1)	Highest (~2500:1 - 6000:1+)
Response Time	Fastest (~1ms)	Moderate (~1-5ms)	Slowest (~4ms+)
Color Accuracy	Poorest	Best	Good
Cost	Lowest	Highest	Mid-range

Even with advancements in liquid crystal technology, LCDs have some inherent drawbacks. Because the crystals and polarizers aren't perfectly light blocking when turned off, light from the backlight can bleed through resulting in poor contrast ratios as they aren't able to produce a pure black.



To get around this, some higher end LCDs now use backlights with local dimming zones. This is essentially just turning off small areas of the backlight that aren't being used. It's not perfect as there's usually some light bleed from areas of the backlight that are being illuminated. Mini LED displays take this one step further by using astonishingly small (50 to 200 micrometers) LEDs for their backlights allowing even smaller local dimming zones.



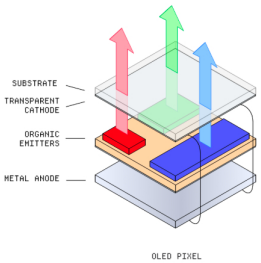


LCDs suffer from somewhat slow response times, as the crystals take some time to change orientation. The alignment of the liquid crystals is optimized for viewing directly in front of the screen resulting in some color and brightness shifts when viewed off axis.

But for all their drawbacks, LCDs are bright, reliable, and most importantly cheap.

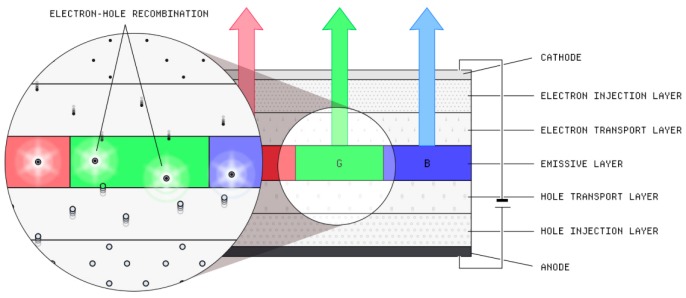
How an OLED works

Organic Light Emitting Diode displays work in a completely different way from LCDs, in fact they are conceptually closer to a mini CRT. As mentioned above, they are self-emissive with each pixel generating its own light, making them much simpler than an LCD with its complicated stack of light conditioning layers.



In its simplest form, an OLED sub-pixel consists of an organic compound sandwiched between an anode and a cathode. When we apply a voltage, current flows through the organic compound which causes it to emit photons. We use different types of organic materials to produce the red, green and blue wavelengths of light we need.

To be more precise though, the light is released as part of a process called electron-hole recombination. This is basically when an electron, which has a negative charge, fills a positively charged 'hole'. A hole is essentially the absence of an electron and usually happens when an electron gains enough energy to jump out of its lattice, leaving behind a sort of empty space.



The bottom of the OLED stack consists of layers to inject and transport these holes up to the

emissive layer while the top consists of layers to inject and transport electrons down to the emissive layer, where they are recombined. When this recombination happens, the wavelength of light is determined by the energy gap of the material used in the emissive layer.

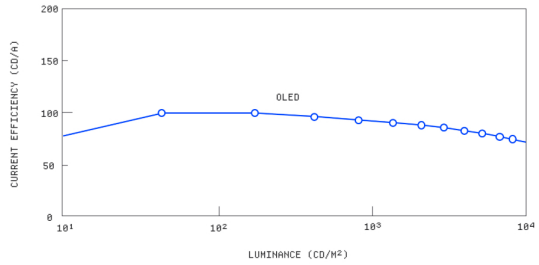
The substrate is sprayed with millions of tiny dots of these organic compounds, allowing them to be placed incredibly precisely. Organic, in this context, just means that these compounds are carbon based.

Because each sub-pixel emits its own light, they can simply be turned off when not in use. This means they can be perfectly black, which an LCD with its backlight bleeding struggles to achieve, while also having the benefit of reducing power consumption in practice.

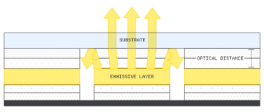
Aside from the excellent contrast and power efficiency, they have a ton of advantages over LCDs. Each pixel can be switched on and off in the sub-millisecond range which makes them perfect as high-refresh rate displays. Due to their lack of polarizers which create inherently directional light, the image quality is largely unaffected by viewing angles and their color accuracy is extremely good. Lastly, because they are much simpler they can be extremely thin and light, even being manufactured on flexible substrates for folding phones.

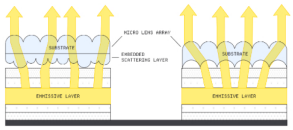
Most of their drawbacks stem from the organic compounds used. Because they are organic, they have a limited lifespan and slowly degrade after use, especially at high brightnesses. This is what causes burn in on OLED displays that are used to display static images for long periods of time, with the blue sub-pixels tending to degrade faster than the others.

They also struggle with brightness - a constraint of the organic materials used - and are less efficient at the higher voltages required to produce brighter light. Most of the advancement in OLED tech has been to address the brightness issue which conveniently also extends the lifespan by allowing them to run at a lower voltage.



Improving brightness is about improving something called Light Extraction Efficiency (LEE) which is how much of the light generated by the emissive layer actually makes it out of the front of the panel. Turns out, even though the distances are small, there is fair amount of reflection that happens inside the stack, causing light to be bounced around internally. To get around this, manufactures use tiny lenses or scattering layers to focus the light and extract it out.





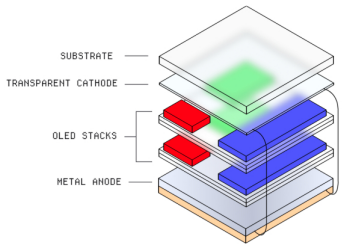
The other approach manufacturers have taken to improve brightness and lifespan, is to combine a single color OLED emissive layer with quantum dot filters to color the light, the same way some LCD panels do. For some reason they tend to use blue, even though I thought that would degrade faster, and that blue light passes through red and green quantum dots. ⁂

All these advancements have made OLEDs pretty competitive with LCDs for most applications, but they still aren't perfect.

The next generation

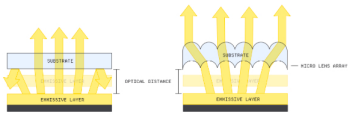
As each technology slowly chips away at their own inherent weaknesses, they are converging on almost the same type of display but it seems as if LCDs are closer to their limit, just trying to make smaller and smaller local dimming zones. There are two really fascinating emerging contenders for the next generation of display: Tandem OLED and MicroLED.

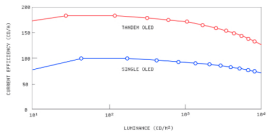
A Tandem OLED, is exactly what it sounds like: two OLED panels glued together. By having two emissive layers stacked on top of each other you can increase the brightness of whole panel while also having the advantage of being able to run each layer at a lower voltage, increasing efficiency and lifespan.



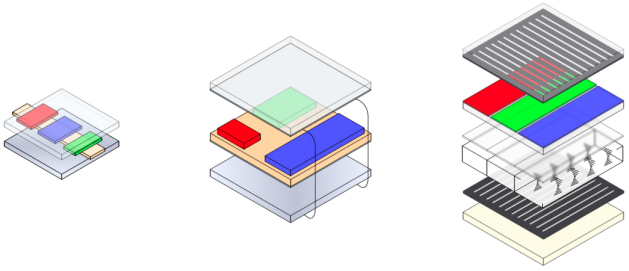
TANDEM OLED PIXEL

They present some new challenges though, in that having two layers increases the optical distance between the bottom layer and the top substrate, making the reflection issue worse. Luckily, those micro lens arrays we talked about earlier help alleviate that and improve the Light Extraction Efficiency. They are, quite understandably, still difficult and expensive to produce though. Apple has used them for the latest generation iPad but it will be interesting to see when they can be scaled up for larger displays.





A MicroLED display probably works the way you expected all screens to work before you started reading this chapter (about 3000 words ago). They consist of absolutely microscopic, in the micrometer range, individual LEDs for each sub-pixel. Because they are inorganic, they aren't prone to lifespan and burn in issues that plague OLEDs and can be much brighter.


MICRO LED PIXEL OLED PIXEL LCD PIXEL

With OLEDs, we can sort of spray the organic compounds onto the substrate, but because these are just tiny LEDs with their own circuitry, they have to be placed individually onto the substrate. This makes them incredibly expensive and difficult to produce for now, with current panels not quite achieving the pixel density required for anything other than the size of display you'd see at a concert or sporting event where viewing distances make up for the gaps between pixels being noticeable.

So in case you thought the modern display was 'solved', it really isn't. It'll be really fascinating to see what happens in the next decade, with even more esoteric technologies emerging all the time.

So what was the point of all of this? Why put all this effort in a book about software, explaining in excruciating detail how a piece of hardware works?

Well, firstly, I enjoy it. And secondly, understanding how a screen works unlocks so much more understanding of how color works on digital displays which is going to feature a lot in the next few chapters. If you don't understand the various trade-offs in display tech, the constraints of digital color seem sort of arbitrary. Not to mention later when we talk about things like rasterisation, GPUs and, shaders.

Anyway, that's it. If you made it to this point without losing the will to live, congrats, you have the same type of mental illness that I do and you are in the exact right place.

--- END ---