**Notes from Interim, Tuesday Oct. 11**

Action Items marked AI

**Tuesday am**

Attendees - Introductions – Stephan Wenger, John Leslie, Mark Duckworth, Roni Evan, Mary Barnes, Paul K, Gonzalo, Jonathan, Steve Botzko, Spencer Dawkins, Andy Pepperell, Brian Baldino, Hadriel Kaplan, Bo, .. see roster

**Note Well**

**Agenda, see slides**

**Use Cases**

Not enough review

**AI**: Roni --Data collaboration – what can do over RTP is ok. Maybe add something to text can use any RTP payload. See what’s happening in RTCWEB on same topic.

Mary - Won’t send use cases to publication because we might want to add something

Gonzalo – Send drafts as a bundle when ready

**Requirement**s

Ready to be reviewed

How comprehensive does it need to be? Does not need to map back everything in the solution

**AI: draft authors, numbered bullets**

1. Necessary to add requirement for updating properties . The term capabilities is problematic here. Add this into requirements. How it should be worded needs to be worked out
2. Adding a third dimension. Needs more discussion on the list. How it should be worded. Reqmt 1b not the right place for it. 1b is about ordering

What does depth mean in terms of images?

1. Multi-view—should be in reqmts or why is it in the document. Also in the use-cases. Add a requirement for this. Roni will draft

Are the requirements for the basic functionality or also the (possible) extensions?

Should we mark some requirement as not for immediate specification. We should document ideas that are for later.

There are 2 types of multi-view- the kind described in the use cases. And there is also 3 –D. the requirement can reference the specific use case. Multi-view is not a good term because it has multiple meanings.

We need a definition of multi-view. Want 2 cameras looking at the same object.

3-D further out. Not included now.

1. Should we add Jonathan’s case – wants to permanently view another endpoint “pin”.

Mark – use case 3.3 mentions switching video and mentions manual control.

**AI: Jonathan** Should add to the use case. Then drive requirement from there. Jonathan will send to mailing list.

**Framework Discussion**

**Mark Duckworth presentation of New Stuff in draft see slides**

More detailed spatial information about media captures

Additional topics –VAD, media source selection, composition and switching algs – switched and composition how to better describe, selecting algs. Between EP and middle box

**Brian present Area and Point of Capture**

Assign coordinate system to a room, independent of any camera and microphone placement.

X coord. Left to right, z low to high, y front to back

Camera right and camera left. Front is closer to the camera

Where is the origin? Brian says it’s the implementer’s choice, it doesn’t matter

No point in describing parts of the room that are irrelevant, but nothing prevents it

In the diagram, example with field of view 1, 0,0 is bottom right hand corner

Area of capture the segment the camera captures

Begin and end for each of the 3 coords x, y, z

What is the info in the second 2 coords useful for? How will they be used?

Stephan- very useful. May want to render a gap, depending on your application.

Make it be able to turn, allowing for curve, geometric correction is possible. This is not new or unused technology.

Roni- the gap between the cameras.. if someone stands between the 2 cameras, it should look good. Current systems do that.

Brian – enables you to get in there with a ruler, may not require this.

Does this degrade gracefully? Allows you to be as accurate as you want to be. But all systems don’t require this level of accuracy.

If you have a curved table, systems don’t care about representing a curved table. Even for 2 rows of curved tables, won’t care.

Want to enable detail, but not require it.

Bevel, actual size. We haven’t drilled down how relates to CLUE’s charter and interoperability.

Roni – renderer can render what wants. But CLUE provides info for the best quality.

Jonathan- units are arbitrary or mm? can be either

This example is real world millimeters

What is optional and what is mandatory? Who really knows this info? people who set it up. Not the people who write software. Can’t make it mandatory. Systems won’t have this info accurately.

We should be fail-soft - as information is removed, less info sent, should still have good info, just not as accurate

Roni- looking at high quality system interoperability.

Purpose of Origin – for preserving ordering

--

**Discussion of Composition and switching algs**. Raise questions not necessarily settle now

Booleans for switched and composed. Is it enough or not? People have suggested more info needs to be conveyed.

Roni has questions about this in the preso he will give

Jonathan – the concept that the media capture is switched, is it the right concept? It’s more that you are receiving a particular capture. Doesn’t think that switched is the right way to describe it. Switching between captures. What is implied lower down in protocol is different than this. Capture with different sources that are switched. Is it one capture where the source is switched, or are you switching between captures? How does this relate to RTP?

**Roni presentation on major issues**

**Capture area**. The real scene is curved

Concerned about the gap between monitors when the captures are displayed. Wants to have information so this is seamless

Where is the point of origin? Current draft not clear on this

What would be a good way to describe what you want?

Jonathan – doesn’t matter where you put the origin as long as you support negative numbers. Only need to know the relative position of the cameras. But don’t need the origin point.

Roni - Importance of knowing the gap between video captures. Where gaps depends on vendor. renderer can make better experience if you know

Doesn’t think millimeters should be a Boolean

**Architectural model** –

First 2 bullets - Agree with arch in the doc but it should be better explained

**AI – draft authors – clarify draft, as per Roni’s comments**

One issue is no different between capture set and scene. Scene is not necessary.

Andy – thinks EP isn’t useful in the specification, makes sense to talk about it, but it doesn’t add anything to specification

Media stream providers and consumers, EP does both, rather than an element of the model.

Roni – need to define media consumer and provider.

Andy- Can have just audio or video, streaming or recording. EP just happens to do both. Or an EP could be either or the other.

Jonathan- agrees EP not important. For example, presentation isn’t connected to a room.

Discussion of what’s a CLUE endpoint. Something running CLUE.

Roni - Add an SDP param TPcap – what attributes you support. Agree to hold discussion of implementation as a separate conversation

Need to define how to define extensions – agree

**AI – draft authors – clarify how to define extensions**

**Composed captures**  – Roni wants more information than just the image is composed.

Andy- this is a precomposed image. A midllebox would like to know this information so that it won’t further compose the image.

If iimage is precomposed of most active speakers, it would change often, and you wouldn’t want a heavyweight message mechanism, would want something more dynamic like VAD. Never thought of sending info of what is in the composition. Just whether the stream is composed or not.

Roni says it’s in RTCP today. SSRC, CSRC, not adequate.

Andy- Need a new orthogonal dynamic channel. VAD for example.

Whether capture is switched and/or precomposed is static so it is appropriate to describe in capture set

Don’t want to change every time there is a change in speaker for active speaker

Need both channels – fast changing and slower changing

Could see offering different composed streams

This is the current speaker composition, this is the current presentation composition, for example

**AI – draft authors, make changes as suggested**

Document is unclear on second point In the slides

Document needs to cover the third point. Should be added to framework, and maybe not mandatory

Relevant to endpoint and to MCU. Associate a media capture with endpoint

4th point – okay

**Tuesday afternoon**

Updated agenda

**Marshall presentation 3-D locations: a coordinate system for telepresence**

Slide with 3 screens, 3 co-located cameras, Axis of symmetry of the particular unit. This gives an origin and a coordinate system.

Question – what has an axis of symmetry? Answer each system

Wants polar coordinates for what cameras are seeing

What doesn’t fit in this model? A person with a podium with camera off to the side.

If there is a podium? It breaks the symmetry

There is math, even in the simple cases, not clear we have to specify the origin. What is gained?

Origin is useful only to know where things you care about are

If it’s implementers choice, doesn’t make anything harder

A podium cannot be rendered in this particular set up.

Similarly, additional capture points inside the room

Origin important for polar cords, not so much for Cartesian

Axis of symmetry- everyone has a middle..

Multiple cameras and screens for different sitting arrangements

We need to discuss multi-row further

The finest grained info- Location of the camera and the direction its pointing and its fov

Should not be a function of the physical walls of the room

Andy- doesn’t replace the capture region, Cartesian positions of the cameras

Paul – preferred origin, doesn’t believe that people will agree where it is

Steve thinks it make sense - Where rows are, as distance from the camera.

Allyn questions ability to specify rows within one camera view

Allyn – meta issue - understands Marshall’s description is correct, but is it usable in this context?

Same comment for defining origin, feels not necessary for the renderer

Hadriel – thinks don’t need to have this info to match mics to videos. Thinks can get that info specifically without having to do analysis to determine it.

Jonathan- this way of describing relative captures breaks down for synthetic captures, such as presentations. He wants a way to describe those. And wants to be able to use the same language for synthetic and real.

Marshall’s only works for physical-based captures. Jonathan doesn’t want 2 languages one for synthetic and another for physical

MCU decides who goes on which screen.. or sharing right and left monitor. This language doesn’t make sense.

Language that makes sense is 2 dimensional. Start with simple case,which doesn’t do trigonometry or geometric transforms

Gonzalo – what about zooming cameras? What about focal length?

Marshall doesn’t like this aspect of the framework. Thinks won’t work well for this case

What about when the view is changing all the time?

Stephen B. – relative adjacency wasn’t sufficient to match audio and video

Mark – It is possible to represent a curved plane as flat. The renderer needs to know same info in both cases. The 1 dimensional description will work for synthetic captures

Andy- wants to say the same thing. Wants to have 0-99 mandatory absolutely needed. Needed for segment switching. Essential.

With origin, position, focal length.. optional.

Marshall wants to build a structure that can be used for a long time, so that it will be there later when people need it.

Roni- on synthetic, presentation, and people cases. Agrees can have simple and more info. not enough text about presentation and the mcu case in the draft. Requirements different.

**AI – draft authors – clarify as per Roni’s suggestion**

Jonathan – doesn’t want to have 2 different languages, want the full geo to be an extension of the simpler language…

---

**Andy VAD - A New thing in framework draft**

See slides

Want middle boxes to be able to switch without having to decode audio

Determine active speaker for intra-room segment switching

VAD algs must be consistent

Standardizing algorithm

Don’t want to disadvantage a receiver that has one only audio stream

Details- video linear range system, muddle box receiving multiple captures needs to figure out which is loudest overall and which is loudest within a particular room. Wants to know the energy level associated with that position.

Being able to determine which capture is from the active speaker. How determine which video is active.

What position(s) hold the loudest speaker as well as over all VAD. Consumer can determine which is loudest whether receives all or one pre-mixed.

Tagging active position in a room, audio energy is not the only way – for example, could have buttons on a pad to choose who is active.

Stream configure message specify VAD alg. Provider says which algs, consumer chooses which

Through an RTP header extension

Security of VAD info, is there something that needs to be decoded? Does it need to be encrypted? Traffic analysis could be done.

Marshall suggests talking with Colin Perkins to get his feedback

**AI – draft authors, add to draft as suggested**

Roni – the number of audio and video streams is not one to one

e

Jonathan is currently working on header encryption and energy levels

Jonathan – don’t want anything on the ACs. But for segment switching. Want it on the VC- the energy of the audio for the this VC. What do you want the semantics to be? Rather than a correlation of audio and video, he thinks what want is info about the speaker is in this camera.

Andy – issue with that if have a 2 VC case.

Thinks doesn’t work. MCU receives from many multi-video has to compare audio.

One receiver with many audios

See Jonathan’s draft for carrying energy in the header. Avtext Jonathan sent reference to the list

Brian - Didn’t want to do it in video, might want to flow control it off

Roni- agrees shouldn’t be in the video. Important to associate, but not mix by putting audio in video stream.

**AI – draft authors – fill out VAD**

Paul – asks about effect of different audio levels.

Stephen B thinks there is a bias.

Bo – it’s a real time position stream.

Jonathan and Andy- what about AC0, AC1, AC2 – do they have the same VAD info?

Rather than AC0 VAD just for itself, etc.

But then have a challenge with AC3- that has to combine the 3

Either the same in each or each describes itself

Paul – if replicate info in all of them.

For further discussion

Jonathan – think about. Understand case of getting soundscape right.

Gonzalo – they are setting up an SDP directorate

---

**Messaging – Allyn**

Difference in messaging type between SDP’s offer answer and CLUE’s “publish and subscribe”. One conversation in SDP, 2 in CLUE.

Way to characterize the difference is that offer answer is symmetric, and CLUE’s proposal is for asymmetric

Roni doesn’t like term “capabilities” would rather use the term “proposal”

Gonzalo- think about backwards compatibility when think whether will use SDP or not

Stephan- will have real time updates

Jonathan- advertise all the available scenes. Don’t want a full update. Want partial updates.

Partial state update rules out SDP.

Messages can happen whenever

Roni and Stephan want to talk about the signaling and transport tomorrow. They feel this is an important part of the framework.

Mary suggests they make slides for tomorrow

**List of Framework Issues to discuss- see slides**

1. Point/area of capture - Brian
2. Layout – Mark
3. Source selection - Jonathan
4. Attributes for capture sets (along with those already for media capture) – per the hierarchy each level should have attributes - Roni
5. Describe composed picture, rather than just a flag, so endpoint can decompose and put these

things on several things (i. e., resuse mechanism for spatial coordinates, etc) more like tree structure- Stephan will write up something - Stephan

1. Consumer description is not fleshed out
2. Relations to RTP – Jonathan
3. Definition of multi-view
4. Origin/coordinates
5. Screens vs cameras – Marshall. Does the media producer need to know do/anything with screens?
6. Multiple row use case
7. VAD – input and output

Priority voting – for what to cover, choose 5.

13, 3, 2, 1 plus Stephan’s signaling and transport

**Mary RTCWEB presentation**

**Overview** – doesn’t need to interoperate with CLUE

**Proposal for CLUE wrt RTCWEB**

* Ensure that SDP usage is compatible and consistent to ensure that CLUE and RTCWEB do not define 2 separate ways of doing the same thing
* Evaluate usages of SDP/RTP as framework is being developed
* CLUE needs to consider RTC WEB decisions in terms of handling multi-streams-
  + Multiplex over a single rtp session
  + Or multiple rtp sessions

Stephan says CLUE spec doesn’t need to reference RTCWEB

Hadriel - CLUE should make sure that RTCWEB doesn’t require a symmetric model

Sending data in the media path- rtp extension headers

Browser support audio video only

Data channel will look different, will be RTCWEB specific

TCP over UDP for example

Need congestion

**Covering of Framework Issues**

**Layout – Mark Duckworth**

Switching.. describe a media capture as being switched. Can be from an EP or an MCU. Framework just has a Boolean. No info on an alg. For how is being switched.

Called auto-switched.

Roni – has an issue with this

What info is he looking for? How often switched, differentiation on the side of the provider. Wants the consumer to know the basis of switching, so can choose which alg. It wants.

**AI – authors add in as Roni suggests**

Jonathan – multipoint case. Scene switching, if have different geometries, need to know the concrete scene. Has a lot of implications. Whether have a switched capture. Or switching among captures.

Have all info

If switching, want to tell which scene you are seeing

Means need to know geometry for each room

If MCU switches and you don’t know, you can’t do meaningful adjustments. Rely on what the MCU does for you

Request the current loud speaker.

Partial update for new scene in conversation. Capture 38. At run time gets info, you are now seeing 38.

Spencer – from Booleans to enums. Is this capping innovation? If not on the list, will you know what to do? These are the 34 things you can do.. restrictive

Roni- description of options can use

How do you describe new options

Alot of algs can decide how to switch

Stephen - Separate negotiation of policy from attributes need to know

Want MCU to be able to know whether what it’s receiving is composed

**AI – authors Separate policy management from attributes**

Want to see the current speaker and the current preso

Switched capture

Jonathan – site switching. How do you ask for site switching?

might need other attributes for composition

for switching.. don’t need to describe on the forward path

Andy- should there be attributes of capture set rows?

This isn’t just for multipoint,

Consider 3-1 point to point case for switching

Mark – where is this leading? What’s the use case?

Roni- value in getting info about what the switching alg. Either by the provider who can offer different algs to the consumer.

Meta information about the streams. Getting 3 switched streams, which is which? Depending on policy, which is which?

Tied to RTP mapping

Would a light weight message be fast enough?

Ask for 3 active speaker, get 3 or 2 +1. Sending a dynamic update. But figure out how to do it. Which participant this is.

**Discussion of RTP**

Jonathan- How do you bind SSRC to a virtual capture?

Assumption demux on SSRC or not. Andy is assuming demuxing on something other than **SSRC** – what was jonathan’s issue?

payload, CSRC. Jonathan was assuming demux on SSRC.

Distinction between real and virtual source.

Potential decryption issues in the MCU

Have identified some constraints.. switching and composition needs to be known

Need to know original source, distinction between virtual and real source. Messaging and RTP implications.

If you know the actual source, need to know what it is – roster list … way we bind?

Tricky to act at high speed..

Static pinning- and he becomes loudest speaker, want to get the data only once. Individual mute.

Someone needs to write some text.

This is intertwined with what are talking about for tomorrow

MCU wants to know if composed or original. This is a Boolean.

Auto-switch Boolean and auto-switch with policy. Haven’t totally decided if auto-switch Boolean is okay

**Plans for tomorrow**

Issue 13, signaling and transport, break,

Area capture- Represent things in the simple way. Agree on this. To be able to use the same language.

Source selection