Help CHS in AGA4 (GPUv6/MI355x)

Process of Analysis

The team considered the perspective of an IC1 receiving the tickets listed below and whether that IC1 could execute on them based on the information in the tickets. The goal of the exercise is to have any CHS technician execute a repair with clear instructions.

	Avnish Content			CHS Content			
	Category	**Ticket example**	**Notes**	**Details**	**Actionable by CHS (approved by CHS)**	**Valid runbook?**	**Actions**
1	~~Card Management~~	~~https://jira-sd.mc1.oracleiaas.com/browse/DO-26141272?filter=381507~~ ~~https://jira-sd.mc1.oracleiaas.com/browse/DO-26162982?filter=381507~~	~~Runbook link in description~~	Duplicate of line 4			Both Tickets here are Smartnic wipe unable to find any DHCP offer. – Analyzing under that bucket.
2	Open Problems	https://jira-sd.mc1.oracleiaas.com/browse/DO-2605626	Runbook links in description Runbook : DCO ticket runbook ARO Runbook:Url : runbook	Open Problems tickets can come in a few ways. 1. TRS tickets that do not have a GRT match a. cut by TRS with "Generic open problems" in summary 2. Open problem tickets manually cut by various service teams. a. No real consistency across tickets and format 3. Multiple runbooks in the description of the ticket with no indication of which runbook should be followed. 4. Faults not prioritized 5. No requested physical action in the ticket. 6. No DCSE run book 7. BMP-825375 / DO-2605626 - was cut HoPS Proxy ticket. • Ticket was not cut by TRS, why?	No	No	☐ BMP-825375 / DO-2605626 has active ILOM faults and should have gone to TRS. ☐ Remove dependecny on HoPS proxy tickets (failed in lego action GpuTailnodePreProvisioning) needs a CHS run book

3	Card IP address doesn't match plan	https://jira-sd.mc1.oracleiaas.com/browse/DO-2625147	Instructions in ticket "Hi Team, Please check and verify cross cabling for smartnics . Smartnic FPF253714F4 should not be visible. "	IP connection on smartnics were incorrect. • CHS tech triaged the system to validate which cables were connected incorrectly and how they needed to be swapped. • The work the CHS tech did here should have been done prior to the ticket being cut. • All connection related tickets should include both endpoint port connection before cutting ticket to CHS • The CHS ticket should have said "smartnic cabling incorrect, please swap the cable endpoint of Serial | Elevation | Port with Serial | Elevation | Port and reach out to the requester to verify" • No actionable runbook for CHS • No Direct/ prescriptive physical request • No Published CHS runbook	No	No	☐ Service Team to engage DCSO to create CHS authorized run book for this failure https://confluence.oci.oraclecorp.com/display/DCOps/DCGSE+Engagement+Process

4	SmartNicRot Wipe and open problems
	
	OR
	
	Smartnic wipe unable to find any DHCP offer
		https://jira-sd.mc1.oracleiaas.com/browse/DO-2616171
		https://jira-sd.mc1.oracleiaas.com/browse/DO-2614127
		
		https://jira-sd.mc1.oracleiaas.com/browse/DO-2616398
			Instructions in description
			1. Reseat the smartnics
			2. Check if any card /cable is faulty
			3. Powerdrain the host.
			4. check/replace the wipe cable if needed
			5. Verify the serial connectivity between rot and smartnic
			.
			Runbook : https://devops.oci.oraclecorp.com/runbooks/CRDMGMT/card-management-ops/crdmgmt-check-tor
			https://devops.oci.oraclecorp.com/runbooks/CRDMGMT/card-management-ops/crdmgmt-check-rot-serial
				Smartnic cannot receive DCHP request. Ticket suggest cabling issues.
				
				▪ runbook link is CRDMGMT devops runbook and not CHS approved documentation
				▪ There is physical action plans for CHS in the ticket
				▪ Linked runbook does have physical request, however they are not detailed enough
				  ▪ No Physical endpoints
				  ▪ Generic request (check smartnics, check TOR, etc...)
				▪ CHS should not be doing triage steps that can be automated or performed by centralized teams. (Check ILOM open Problems, check smartnic presence)
				▪ No actionable runbook for CHS
				▪ No Direct/ prescriptive physical request
				▪ No Published CHS runbook
					No	No	☐ Service Team to engage DCSO to create CHS authorized run book for this failure https://confluence.oci.oraclecorp.com/display/DCOps/DCGSE+Engagement+Process

5	Fix Open Problems	https://jira-sd.mc1.oracleiaas.com/browse/DO-2616570 https://jira-sd.mc1.oracleiaas.com/browse/DO-2622675	Is the expectation to follow the GRT for the open problem? CHS docs have FMS failure handling link https://confluence.oraclecorp.com/confluence/display/DCKS/FMA+Code+Handling - does this apply? Typically open problems are listed and CHS is asked to work on them.	Same issues as Row #3 above	Same issues as Row #3 above	Same issues as Row #3 above	Same issues as Row #3 above

6	Cable Replacement	https://jira-sd.mc1.oracleiaas.com/browse/DO-2569753	Tickets assumes CHS can debug link down , it provides reference docs and repair history link
			
			Link is hard down. Please troubleshoot.
			
			You may refer to the following documentation for assistance with link troubleshooting:
			
			• DCO Rack Cabling Standards
			• Troubleshooting AOC Link Connectivity
			• Cable Installation
			• Network and fiber testing
			
			Check the history for aga4-c1-b9-t2-r4 to aga4-c1-t3-c13-r4 DO link tickets here
			
			Open the following grafana link to review link status, light levels, temperature, etc over time. Grafana
			
			aga4-c1-b9-t2-r4 (s/n FGN245100AW) is located in Building aga4, Rack 3535, elevation 6
			aga4-c1-t3-c13-r4 (s/n HBG24470L6S) is located in Building aga4, Rack 3912, elevation 11
			
			Note - the ports have been disabled.
				▪ Too much information unable to determine correct step quickly
				▪ Not clear what steps should be followed and in what order
				▪ Multiple CHS runvbooks provided. none apply
				  ▪ Runbooks provided don't match the connection type
				▪ Contradicting request
				  ▪ Summary says please request cable
				  ▪ Description says "please troubleshoot"
				▪ Connection endpoints given are rarely the direct connection, they are either jump connection or breakout connections.
				▪ CHS should not be asked to perform historical analysis or light level triage
				▪ Eventually the ticket gives the connection points, but this should be the ONLY information provided.
				
				`aga4-c1-b9-t2-r4 (s/n FGN245100AW) is located in Building aga4, Rack 3535, elevation 6 aga4-c1-t3-c13-r4 (s/n HBG24470L6S) is located in Building aga4, Rack 3912, elevation 11`
					No	No	☐ Service Team to engage DCSO to create CHS authorized run book for this failure https://confluence.oci.oraclecorp.com/display/DCOps/DCGSE+Engagement+Process

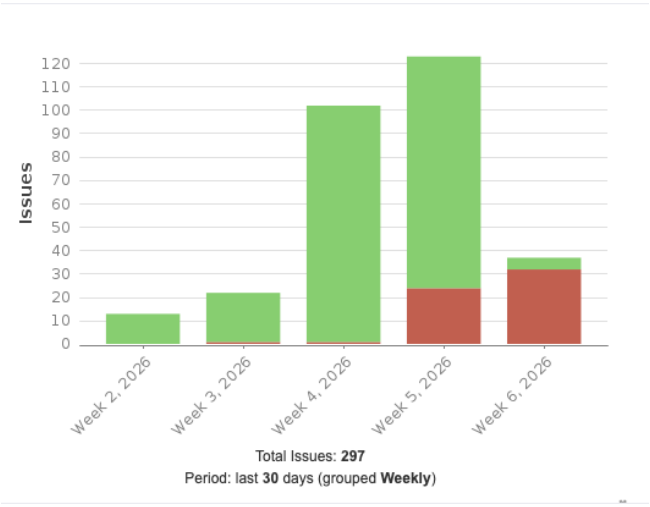
7	Link Down from CPV health check	https://jira-sd.mc1.oracleiaas.com/browse/DO-2616924	Link is identified and CHS is asked to check why it is down. OR all links are down and open problems are listed. Unclear if CHS docs cover link debugging.	▪ RDMA connection ▪ Missing if HN or TN ▪ OS information provided. CHS expected to translate to physical ports ▪ Generic Ticket Summary ▪ component item is set to hardware-gpu when the ticket looks to be link repair and should say network / link ▪ Physical port information not clear (nested in OS output) ▪ Manual ticket that should have been automated with information that CHS expects.	No	No	☐ CPV root cause manual ticket creation ☐ CPV to automate link source and destination physical info ☐ CPV should use TRS/GRT recipes ☐ CPV Team to engage DCSO to create CHS authorized run book for this failure https://confluence.oci.oraclecorp.com/display/DCOps/DCGSE+Engagement+Process ☐

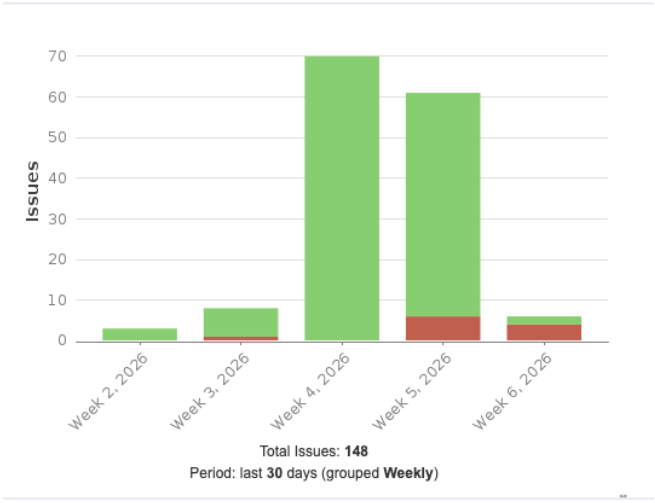
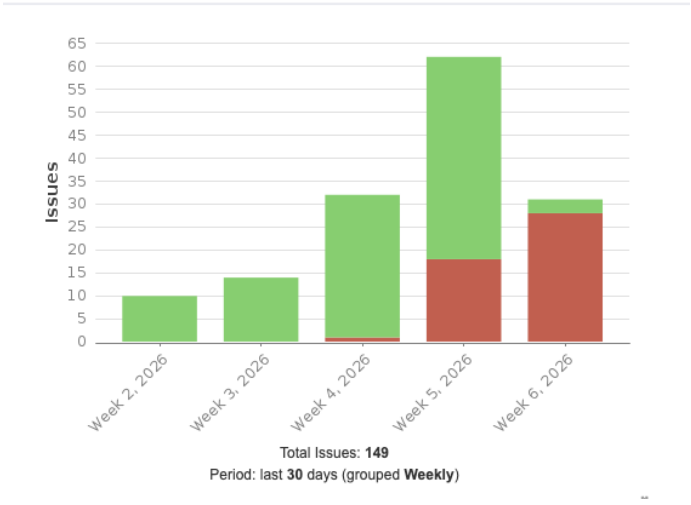
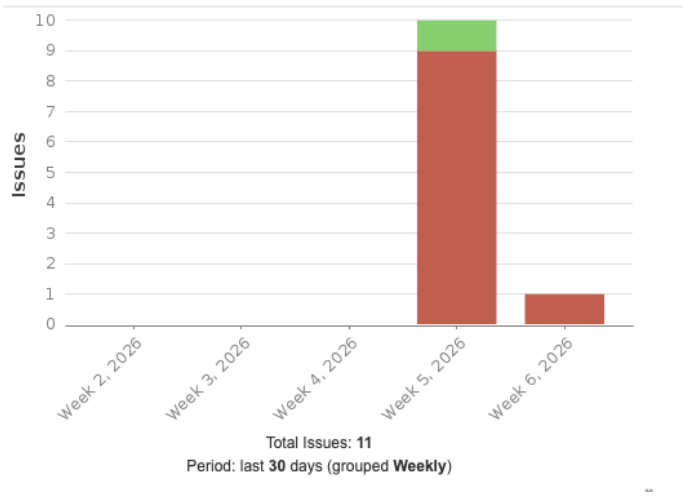
8	Health check v2	https://jira-sd.mc1.oracleiaas.com/browse/DO-2625265	" "issue": "CDFP cable connection mismatch detected. GPU PCI addresses and module IDs do not match expected configuration", "suggestion": "Verify CDFP cable connections between GPUs match the expected mapping. Check physical cable connections and GPU seating", "	▪ Initial request is generic "Health check is reporting to check CDFP cable connection" ▪ What about the connection? Is it down? ▪ Expanded upon later in CPV outputs ▪ No prescriptive physical request ▪ CPV should include expected port mapping ▪ Ticket summary needs to be more prescriptive ▪ CHS is expected to translate ▪ "Please look into this" ▪ "Verify CDFP connections match the expected mapping" ▪ `what CDFP connection? What expected mapping?`	No	No	☐ CPV Team to engage DCSO to create CHS authorized run book for this failure ☐ SCO to verify if factory does PCI mapping validation at L10 testing
9	Power SHelf	https://jira-sd.mc1.oracleiaas.com/browse/DO-2571809	" check RU41 power shelf Right Lite ON. Failed PS1. Appears shelf RU41 has failed"	• Ticket self created by CHS form visual inspection (floor walk) • Power shelves may not be monitored in OC1	No	No	☐ Identify who owns power shelf monitoring? (pulse? DCICM?) ☐ Once owner is identified, automation needs created for these tickets.
10	Check and correct network connection per topology	https://jira-sd.mc1.oracleiaas.com/browse/DO-2600701	Please check below connection and correct it according to below topology. Correct connection: aga4-c1-nb64-t2-r7 <> Ethernet23/1 <==>. aga4-c1-t3-c14-r7 <>. Ethernet21/1	• Host names are provided instead of serials (CHS doesn't work off hostnames) • Summary and description don't match • No CHS approved run book provided • End points do not include full details (breakout) • Request in ticket is not clear - no clear action plan ○ what does "correct" connection mean? Reseat/Replace	No	No	☐ Service Team to engage DCSO to create CHS authorized run book for this failure https://confluence.oci.oraclecorp.com/display/DCOps/DCGSE+Engagement+Process

11	Cable validation failures	https://jira-sd.mc1.oracleiaas.com/browse/DO-2612700	Links that are not cabled to the correct switches (MAC/ipv6 address not found at all): aga4-c1-b15-t0-r4-u1-gpu-expander4, SLED1/PCIE1/1, aga4-q3-b38-t0-r1, Ethernet28/1 is empty. Check the link. aga4-c1-b15-t0-r4-u1-gpu-expander4, SLED1/PCIE2/1, aga4-q3-b38-t0-r2, Ethernet28/1 is empty. Check the link. CHS/DO is authorized by Compute to work on this host in the allocated state in the holding pool with the CPV instance. This is a non-terminating repair. **Compute Product Validation** **Job ID:** c0035613-a4ec-4898-9d29-0a954831bbe9 **Test Name:** cable_validation **Asset ID:** 2539XK10CS **Need help?:** Submit a ticket to the HPC queue	▪ Summary is generic - no indication of "what to do" ▪ Description is not clear ▪ States cables are not cabled to the correct switches, but then suggest cables are "empty" ▪ No CHS approved run book provided ▪ Host names provided instead of physical serial and elevations. ▪ Direct port mapping	No	No	☐ Service Team to engage DCSO to create CHS authorized run book for this failure https://confluence.oci.oraclecorp.com/display/DCOps/DCGSE+Engagement+Process

12	RDMA network connection unstable	https://jira-sd.mc1.oracleiaas.com/browse/DO-2616078	BM.GPU.MI355X.8 2540XK101L is connected to chassis 2543XK105F.
			The following RDMA network interfaces are unstable going UP and DOWN.
			For each network cable serial number confirm there are no other associated tickets.
			Please **replace** the RDMA network cable connection(s) listed below and visually confirm there is link.
			
			If CDFP cabling was disconnected to perform this task, power cycle the host to confirm no open problems are present and all PCIE devices have trained.
			
			After performing this task please resolve this ticket to allow automated testing to resume.
			
			Engage OCIGNOC or CNE for assistance when repeated tickets have been cut for a network cable.
				• Good initial prescriptive request
				• Host names provided instead of physical serial and elevations.
				• Information is over all good, but still confusing
				  ◦ Could be consolidated and better for easier digestion
				• CHS should not have perform ticket history analysis
				  ◦ "For each network cable serial number confirm there are no other associated tickets"
				• Ticket looks to have crossover information for other systems that don't apply here.
				• CHS should be be owner of routing tickets to different after physical action completed
				  ◦ "Engage OCIGNOC or CNE for assistance when repeated tickets have been cut for a network cable."
				• No Approved CHS run-book in the ticket
				• CHS does not have training documentation on TRS/CPV (terminating vs non terminating) and how to interface with these tickets type.
					No	No	☐ Service Team to engage DCSO to create CHS authorized run book for this failure https://confluence.oci.oraclecorp.com/display/DCOps/DCGSE+Engagement+Process
							
							☐ CHS needs trained on TRS service. ticket types and how to work them
							
							☐ CHS needs trianed on CPV service, ticket types, and how to work them

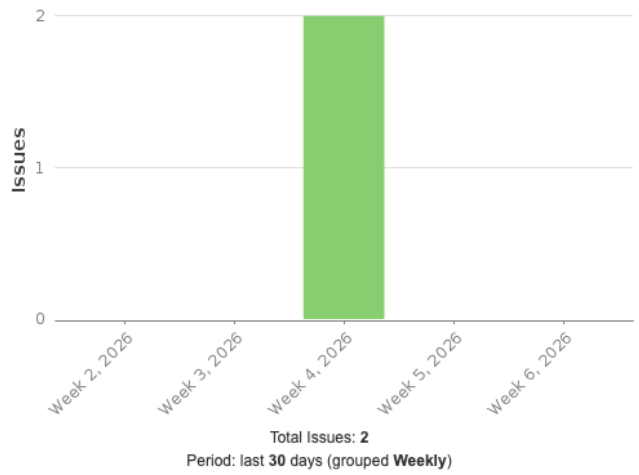
13	Reseat or replace PSU	https://jira-sd.mc1.oracleiaas.com/browse/DO-2599694	**Please reseat** the broken headnode component(s): PSU **DCO Actions** Reseat / Replace PSUs https://confluence.oci.oraclecorp.com/x/11cYzw https://confluence.oraclecorp.com/confluence/x/0IMvJAM **Special Instructions** power supplies are unable to provide sufficient power to the system reseat all PSUs and install a PSU in any locations that are missing one. Replace any PSU reporting as failed	• TRS ticket • Action plan is clear and prescriptive • Description doesn't specify what to do after the action • Indicates to create an RHS ticket if you have problems doesn't align with the TRS process	**Yes**	**Yes**	☐ CHS needs trained on TRS service. ticket types and how to work them ☐ TRS needs to update ticket description to be more clear on how to work the ticket (aka resolve ticket and TRS will validate, etc..)

Overall Open Problems		
• 1052		
Open Problem Categories	**Charts/Filter**	**Analysis**

TRS Cut tickets GENERIC_OPEN_PROBLEMS 827		Tickets are cut with Open problems that are not aligned to GRT Recipes. AI: Identify which organization should receive these issues, during NPI to implement repair recipes, to improve support during Handovers AI: Implement Ticket routing to support GENERIC_OPEN_PROBLEMS for NPI
Human Cut Open Problems 225		AI: Identify authoring team members and inform them on how to cut prescriptive tickets using GRT information. AI: SMG_SRE Tool is used to cut tickets and can implement GRT recipes to improve action items for Repair. This can be used by CPV/HOPS/REPAIR operators AI: Tickets cut by CPV operations can leverage TRS/CPV integration to cut prescriptive ILOM tickets using GRT. AI: Tickets cut by HOPS operations can leverage TRS/HOPS integration to cut prescriptive ILOM tickets using GRT. AI: Spurious Faults tickets being cut should not be cut to CHS. SOP to review if ILOM open problems are spurious before cutting CHS ticket
SmartNicRotWipe		
Non-TRS Cut Human cut Tickets 34		
Reseat or replace PSU		

TRS Cut Tickets

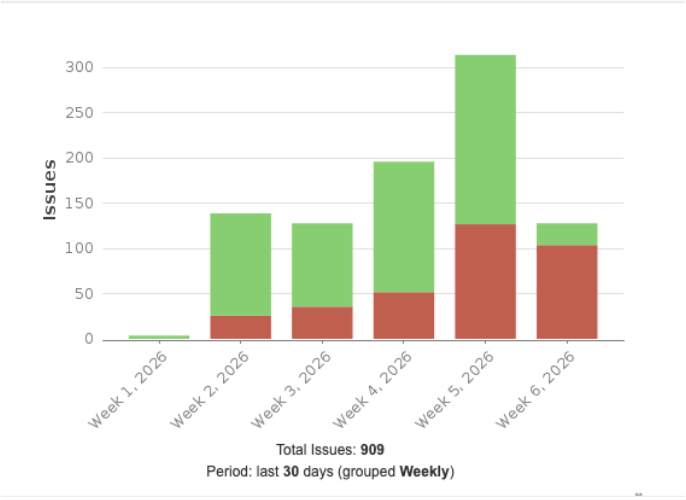
4



Human Cut Tickets to DO

AGA4

17061



Last 30 : 909

- AGA-CPV 145
- LINK_DOWN 69
- Down.. Links 14
- owner:Networking 136
- SMG_SRE_TRIAGE_TOOL 111
- SMARTNIC_WIPE_FAILURE 4
- CPV Job Failure 52
- CPV Job Timeout 11
- Open Problems 71
- OP Found on Host 14
- OP found on TN 11
- Sanitization 22
- RHS Action Plan 27
- Data not found in StoreKeeper 22
- Storekeeper data is missing 2
- Ethernet Cable 8
- Rack Location Unreserved 24
- SUNVTS 4
- SMARTNIC_DHCP_FAILURE_DO_QUEUE 3
- RDMA 34
- SWAP GPU 5
- EMMC_BOOT_REACHED_MAX_RETRY 2
- Networking Misc: 16
- Destruction Bin Audit 4
- Replace GPU 12
- Cable Check 8
- Wipe and rebuild failed on host after multiple retries. Failing HoPS state /job 1
- Fabric Test Failed 2
- SMARTNIC_ESCALATED_RECIPE 6
- Smartnic wipe unable to find any DHCP offer 2
- Misc Smartnic 7
- Power 8
- DIMM 3
- DCS 2
- health_check 2
- miniHPL Test failure 1
- Incorrect Inventory 1